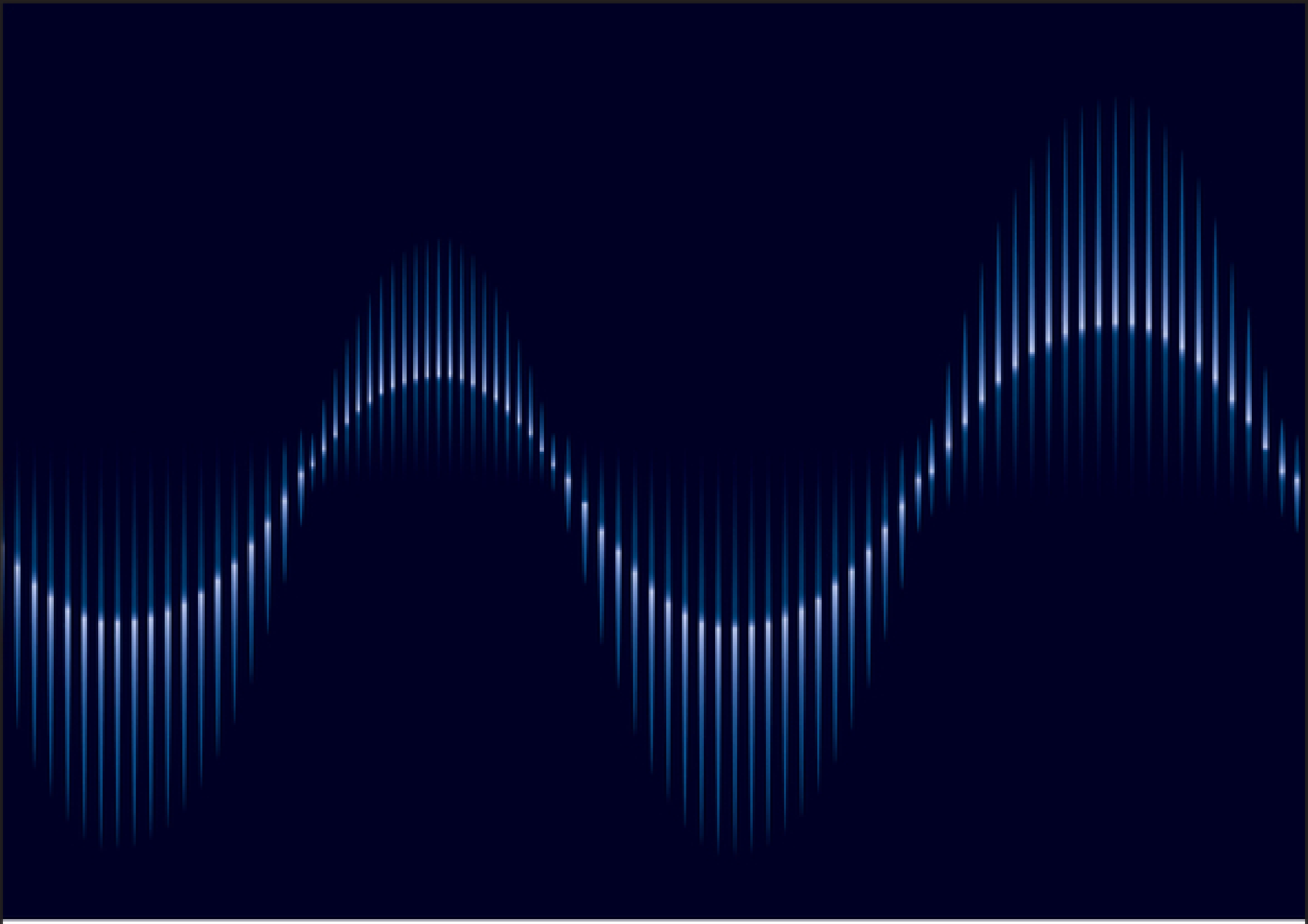


STATISTICS AND PROBABILITY WITH APPLICATIONS FOR ENGINEERS AND SCIENTISTS USING MINITAB, R AND JMP

SECOND EDITION

BHISHAM C. GUPTA | IRWIN GUTTMAN | KALANKA P. JAYALATH

with website

WILEY

STATISTICS AND PROBABILITY WITH APPLICATIONS FOR ENGINEERS AND SCIENTISTS USING MINITAB, R AND JMP

STATISTICS AND PROBABILITY WITH APPLICATIONS FOR ENGINEERS AND SCIENTISTS USING MINITAB, R AND JMP

Second Edition

Bhisham C. Gupta
Professor Emeritus of Statistics
University of Southern Maine
Portland, ME

Irwin Guttman
Professor Emeritus of Statistics
SUNY at Buffalo and
University of Toronto, Canada

Kalanka P. Jayalath
Assistant Professor of Statistics
University of Houston–Clear Lake
Houston, TX

WILEY

This second edition first published 2020

Edition History
John Wiley & Sons, Inc. (1e, 2013)

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Gupta, Bhisham C., 1942- author. | Guttman, Irwin, author. |
Jayalath, Kalanka P., author.
Title: Statistics and probability with applications for engineers and
scientists using Minitab, R and JMP / Bhisham C. Gupta, Irwin Guttman,
Kalanka P. Jayalath.
Other titles: Statistics and probability with applications for engineers
and scientists
Description: Second edition. | Hoboken : Wiley, 2020. | Revision of:
Statistics and probability with applications for engineers and
scientists. 2013. | Includes bibliographical references and index.
Identifiers: LCCN 2019035384 (print) | LCCN 2019035385 (ebook) | ISBN
9781119516637 (cloth) | ISBN 9781119516644 (adobe pdf) | ISBN
9781119516620 (epub)
Subjects: LCSH: Probabilities. | Mathematical statistics.
Classification: LCC QA273 .G85 2020 (print) | LCC QA273 (ebook) | DDC
519.5–dc23
LC record available at https://lccn.loc.gov/2019035384
LC ebook record available at https://lccn.loc.gov/2019035385

Cover Design: Wiley
Cover Image: © ioat/Shutterstock

Set in 10/12pt Computer Modern by SPi Global, Chennai, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

In the loving memory of my parents, Roshan Lal and Sodhan Devi
-Bhisham

In the loving memory of my parents, Anna and Samuel Guttman
-Irwin

In the loving memory of my parents, Premadasa Jayalath and Chandra Unanthanna
-Kalanka

Contents

PART I FUNDAMENTALS OF PROBABILITY AND STATISTICS

PART II STATISTICS IN ACTIONS

Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue.

R. A. Fisher

Preface

AUDIENCE

This is an introductory textbook in applied statistics and probability for undergraduate students in engineering and the natural sciences. It begins at a level suitable for those with no previous exposure to probability and statistics and carries the reader through to a level of proficiency in various techniques of statistics. This text is divided into two parts: Part I discusses descriptive statistics, concepts of probability, probability distributions, sampling distributions, estimation, and testing of hypotheses, and Part II discusses various topics of applied statistics, including some reliability theory, data mining, cluster analysis, some nonparametric techniques, categorical data analysis, simple and multiple linear regression analysis, design and analysis of variance with emphasis on 2^k factorial designs, response surface methodology, and statistical quality control charts of phase I and phase II.

This text is suitable for a one- or two-semester undergraduate course sequence. The presentation of material gives instructors a lot of flexibility to pick and choose topics they feel should make up the coverage of material for their courses. However, we feel that in the first course for engineers and science majors, one may cover Chapter 1 and 2, a brief discussion of probability in Chapter 3, selected discrete and continuous distributions from Chapter 4 and 5 with more emphasis on normal distribution, Chapter 7–9, and couple of topics from Part II that meet the needs and interests of the particular group of students. For example, some discussion of the material on regression analysis and design of experiments in Chapter 15 and 17 may serve well. Chapter 11 and 12 may be adequate to motivate students' interest in data science and data analytics. A two-semester course may cover the entire book. The only prerequisite is a first course in calculus, which all engineering and science students are required to take. Because of space considerations, some proofs and derivations, certain advanced level topics of interest, including Chapter 20 and 21 on statistical quality control charts of phase I and phase II, are not included in the text but are available for download via the book's website: www.wiley.com/college/gupta/statistics2e.

MOTIVATION

Students encounter data-analysis problems in many areas of engineering or natural science curricula. Engineers and scientists in their professional lives often encounter situations requiring analysis of data arising from their areas of practice. Very often, they have to plan the investigation that generates data (an activity euphemistically called the design of experiments), analyzes the data obtained, and interprets the results. Other problems and investigations may pertain to the maintenance of quality of existing products or the development of new products or to a desired outcome in an investigation of the underlying mechanisms governing a certain process. Knowing how to "design" a particular investigation to obtain reliable data must be coupled with knowledge of descriptive and inferential statistical tools to analyze properly and interpret such data. The intent of this textbook is to expose the uninitiated to statistical methods that deal with the generation of data for different (but frequently met) types of investigations and to discuss how to analyze and interpret the generated data.

HISTORY

This text has its roots in the three editions of Introductory Engineering Statistics, first co-authored by Irwin Guttman and the late, great Samuel Wilks. Professor J. Stuart Hunter (Princeton University), one of the finest expositors in the statistics profession, a noted researcher, and a colleague of Professor Wilks, joined Professor Guttman to produce editions two and three. All editions were published by John Wiley & Sons, with the third edition appearing in 1982. The first edition of the current text was published in 2013.

APPROACH

In this text, we emphasize both descriptive and inferential statistics. We first give details of descriptive statistics and then continue with an elementary discussion of the fundamentals of probability theory underlying many of the statistical techniques discussed in this text. We next cover a wide range of statistical techniques such as statistical estimation, regression methods, nonparametric methods, elements of reliability theory, statistical quality control (with emphasis on phase I and phase II control charts), and process capability indices, and the like. A feature of these discussions is that all statistical concepts are supported by a large number of examples using data encountered in real-life situations. We also illustrate how the statistical packages MINITAB® Version 18, R® Version 3.5.1, and JMP® Version 9, may be used to aid in the analysis of various data sets.

Another feature of this text is the coverage at an adequate and understandable level of the design of experiments. This includes a discussion of randomized block designs, one- and two-way designs, Latin square designs, 2^k factorial designs, response surface designs, among others. The latest version of this text covers materials on supervised and unsupervised learning techniques used in data mining and cluster analysis with a great exposure in statistical computing using R software and MINITAB. As previously indicated, all this is illustrated with real-life situations and accompanying data sets, supported by MINITAB, R, and JMP. We know of no other book in the market that covers all these software packages.

WHAT IS NEW IN THIS EDITION

After a careful investigation of the current technological advancement in statistical software and related applications as well as the feedback received from the current users of the text, we have successfully incorporated many changes in this new edition.

- R software exhibits along with their R code are included.
- Additional R software help for beginners is included in Appendix D.
- MINITAB software instructions and contents are updated to its latest edition.
- JMP software instructions and contents are updated to its latest edition.
- New chapters on Data Mining and Cluster analysis are included.

- An improved chapter on Response Surface Design has brought back to the printed copy from the book website.
- The *p*-value approach is emphasized, and related practical interpretations are included.
- The visibility of the theorems and definitions are improved and well formatted.
- Graphical exhibits are provided to improve the visualizations.

HALLMARK FEATURES

Software Integration

As previously indicated, we incorporate MINITAB and R throughout the text and complete R exhibits with their outputs (Appendix D) and associated JMP exhibits are available on the book's website: www.wiley.com/college/gupta/statistics2e. Our step-by-step approach to the use of the software packages means no prior knowledge of their use is required. After completing a course that uses this text, students will be able to use these software packages to analyze statistical data in their fields of interest.

Breadth of Coverage

Besides the coverage of many popular statistical techniques, we include discussion of certain aspects of sampling distributions, nonparametric tests, reliability theory, data mining, cluster analysis, analysis of categorical data, simple and multiple linear regression, design of experiments, response surface methodology, and phase I and phase II control charts.

Design of experiments, response surface methodology, regression analysis are treated in sufficient breadth and depth to be appropriate for a two-course sequence in engineering statistics that includes probability and the design of experiments.

Real data in examples and homework problems illustrate the importance of statistics and probability as a tool for engineers and scientists in their professional lives. All the data sets with 20 or more data points are available on the website in three formats: MINITAB, Microsoft Excel, and JMP.

Case studies in most chapters further illustrate the importance of statistical techniques in professional practice.

STUDENT RESOURCES

Data sets for all examples and homework exercises from the text are available to students on the website in MINITAB, Microsoft Excel, and JMP format. The sample data sets were generated using well-known statistical sampling procedures, ensuring that we are dealing with random samples. An inkling of what this may entail is given throughout the text (see, for example, Section 7.1.2). The field of sampling is an active topic among research statisticians and practitioners, and references to sampling techniques are widely available in books and journal articles. Some of these references are included in the bibliography section.

Other resources on the book website www.wiley.com/college/gupta/statistics2e available for download include:

Solutions Manual to all odd numbered homework exercises in the text.

INSTRUCTOR RESOURCES

The following resources are available to adopting instructors on the textbook website: www.wiley.com/college/gupta/statistics2e.

Solutions Manual to all homework exercises in the text.
Lecture slides to aid instructors preparing for lectures.
Data sets for all examples and homework exercises from the book, in three formats: Minitab, Microsoft Excel, and JMP.

Errata We have thoroughly reviewed the text to make sure it is as error-free as possible. However, any errors discovered will be listed on the textbook website. If you encounter any errors as you are using the book, please send them directly to the authors bcgupta@maine.edu, so that the errors can be corrected in a timely manner on the website, and for future editions. We also welcome any suggestions for improvement you may have, and thank you in advance for helping us improve the book for future readers.

Acknowledgments

We are grateful to the following reviewers and colleagues whose comments and suggestions were invaluable in improving the text:

Zaid Abdo, University of Idaho
Erin Baker, University of Massachusetts
Bob Barnet, University of Wisconsin-Platteville
Raj Chhikara, University of Houston, Clear Lake
Prem Goel, Ohio State University
Boetticher, Gary D, University of Houston, Clear Lake
Mark Gebert, University of Kentucky
Subir Ghosh, University of California, Riverside
Ramesh Gupta, University of Maine
Rameshwar Gupta, University of New Brunswick, Canada
Xiaochun Jiang, North Carolina Agricultural and Technical State University
Dennis Johnston, Baylor University
Gerald Keller, Joseph L. Rotman School of Management, University of Toronto
Kyungduk Ko, Boise State University
Paul Kvam, Georgia Institute of Technology
Bin Li. Louisiana State University
Thunshun Liao, Louisiana State University
Jye-Chyi Lu, Georgia Institute of Technology
Sumona Mondal, Clarkson University
Janbiao Pan, California Poly State University
Anastassios Perakis, University of Michigan
David Powers, Clarkson University
Ali Touran, Northeastern University
Leigh Williams, Virginia Polytechnic and State University
Tian Zheng, Columbia University
Jingyi Zhu, University of Utah

We thank William Belcher, Darwin Davis, Julie Ellis, Pushpa Gupta, Mohamad Ibourk, James Lucas, Mary McShane-Vaughn, Louis Neveux, and Phil Ramsey who helped find suitable data sets for the case studies. We also thank Laurie McDermott for her help in typing some parts of this manuscript. Special thanks are due to Eric Laflamme for helping write JMP/Excel procedures and creating PowerPoint presentations, George Bernier for helping write Excel work books and macros, and Patricia Miller and Brenda Townsend for editing Power Point Slides and some parts of the manuscript. We appreciate Terry Scott and Chamila Meetiyagoda for reviewing R codes and new edition of the manuscript.

We acknowledge Minitab Inc. and SAS Institute Inc. for permitting us to print MINITAB and JMP screenshots in this book. We also acknowledge the R Core team for allowing us to use open access R software.

Portions of the text are reproduced with permission from the American Society for Quality (ASQ), Applied Statistics for the Six Sigma Green Belt and Statistical Quality Control for the Six Sigma Green Belt by Gupta and Fred Walker (2005, 2007).

We would also like to express our thanks and appreciation to the individuals at John Wiley, for their support, confidence, and guidance as we have worked together to develop this project.

The authors would like to gratefully thank their families. Bhisham acknowledges the patience and support of his wife, Swarn; daughters, Anita and Anjali; son, Shiva; sons-in-law, Prajay and Mark; daughter-in-law, Aditi; and wonderful grandchildren, Priya, Kaviya, Ayush, Amari, Sanvi, Avni and Dylan. For their patience and support, Irwin is grateful to his wife, Mary; son, Daniel; daughters, Karen and Shaun; wonderful grandchildren, Liam, Teia, and Sebastian; brothers and their better halves, Alvin and Rita, and Stanley and Gloria. Kalanka appreciates the support of his wife, Chamila; daughters, Nesandi and Minudi.

BHISHAM GUPTA
IRWIN GUTTMAN
KALANKA JAYALATH

About The Companion Site

This book is accompanied by a companion website:

www.wiley.com/college/gupta/statistics2e

The website includes materials for students and instructors:

Instructors

Chapters 20 and 21
Data sets
PowerPoint presentations
Complete solutions manual
Certain proofs and derivations
Some statistical tables
JMP files
R exhibits

Students

Chapters 20 and 21
Data sets
Partial solutions Manual
Certain proofs and derivations
Some statistical tables
JMP files
R exhibits

Chapter 1

INTRODUCTION

Statistics, the discipline, is the study of the scientific method. In pursuing this discipline, statisticians have developed a set of techniques that are extensively used to solve problems in any field of scientific endeavor, such as in the engineering sciences, biological sciences, and the chemical, pharmaceutical, and social sciences.

This book is concerned with discussing these techniques and their applications for certain experimental situations. It begins at a level suitable for those with no previous exposure to probability and statistics and carries the reader through to a level of proficiency in various techniques of statistics.

In all scientific areas, whether engineering, biological sciences, medicine, chemical, pharmaceutical, or social sciences, scientists are inevitably confronted with problems that need to be investigated. Consider some examples:

- An engineer wants to determine the role of an electronic component needed to detect the malfunction of the engine of a plane.
- A biologist wants to study various aspects of wildlife, the origin of a disease, or the genetic aspects of a wild animal.
- A medical researcher is interested in determining the cause of a certain type of cancer.
- A manufacturer of lenses wants to study the quality of the finishing on intraocular lenses.
- A chemist is interested in determining the effect of a catalyst in the production of low-density polyethylene.
- A pharmaceutical company is interested in developing a vaccination for swine flu.
- A social scientist is interested in exploring a particular aspect of human society.

In all of the examples, the first and foremost work is to define clearly the objective of the study and precisely formulate the problem. The next important step is to gather information to help determine what key factors are affecting the problem. Remember that to determine these factors successfully, you should understand not merely statistical methodology but relevant nonstatistical knowledge as well. Once the problem is formulated and the key factors of the problem are identified, the next step is to collect the

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

data. There are various methods of data collecting. Four basic methods of statistical data collecting are as follows:

- A designed experiment
- A survey
- An observational study
- A set of historical data, that is, data collected by an organization or an individual in an earlier study

1.1 DESIGNED EXPERIMENT

We discuss the concept of a designed experiment with an example, “Development of Screening Facility for Storm Water Overflows” (taken from Box et al., 1978, and used with permission). The example illustrates how a sequence of experiments can enable scientists to gain knowledge of the various *important factors* affecting the problem and give insight into the objectives of the investigation. It also indicates how unexpected features of the problem can become dominant, and how experimental difficulties can occur so that certain planned experiments cannot be run at all. Most of all, this example shows the importance of common sense in the conduct of any experimental investigation. The reader may rightly conclude from this example that the course of a real investigation, like that of true love, seldom runs smoothly, although the eventual outcome may be satisfactory.

1.1.1 Motivation for the Study

During heavy rainstorms, the total flow coming to a sewage treatment plant may exceed its capacity, making it necessary to bypass the excess flow around the treatment plant, as shown in Figure 1.1.1a. Unfortunately, the storm overflow of untreated sewage causes pollution of the receiving body of water. A possible alternative, sketched in Figure 1.1.1b, is to screen most of the solids out of the overflow in some way and return them to the plant for treatment. Only the less objectionable screened overflow is discharged directly to the river.

To determine whether it was economical to construct and operate such a screening facility, the Federal Water Pollution Control Administration of the Department of the Interior sponsored a research project at the Sullivan Gulch pump station in Portland, Oregon. Usually, the flow to the pump station was 20 million gallons per day (mgd), but during a storm, the flow could exceed 50 mgd.

Figure 1.1.2a shows the original version of the experimental screening unit, which could handle approximately 1000 gallons per minute (gpm). Figure 1.1.2a is a perspective view, and Figure 1.1.2b is a simplified schematic diagram. A single unit was about seven ft high and seven ft in diameter. The flow of raw sewage struck a rotating collar screen at a velocity of five to 15 ft/s. This speed was a function of the flow rate into the unit and hence a function of the diameter of the influent pipe. Depending on the speed of the rotation of this screen and its fineness, up to 90% of the feed penetrated the collar screen. The rest of the feed dropped to the horizontal screen, which vibrated to remove excess water. The solids concentrate, which passed through neither screen, was sent to the sewage treatment plant. Unfortunately, during operation, the screens became clogged with solid matter, not only sewage but also oil, paint, and fish-packing wastes. Backwash sprays were therefore installed for both screens to permit cleaning during operation.

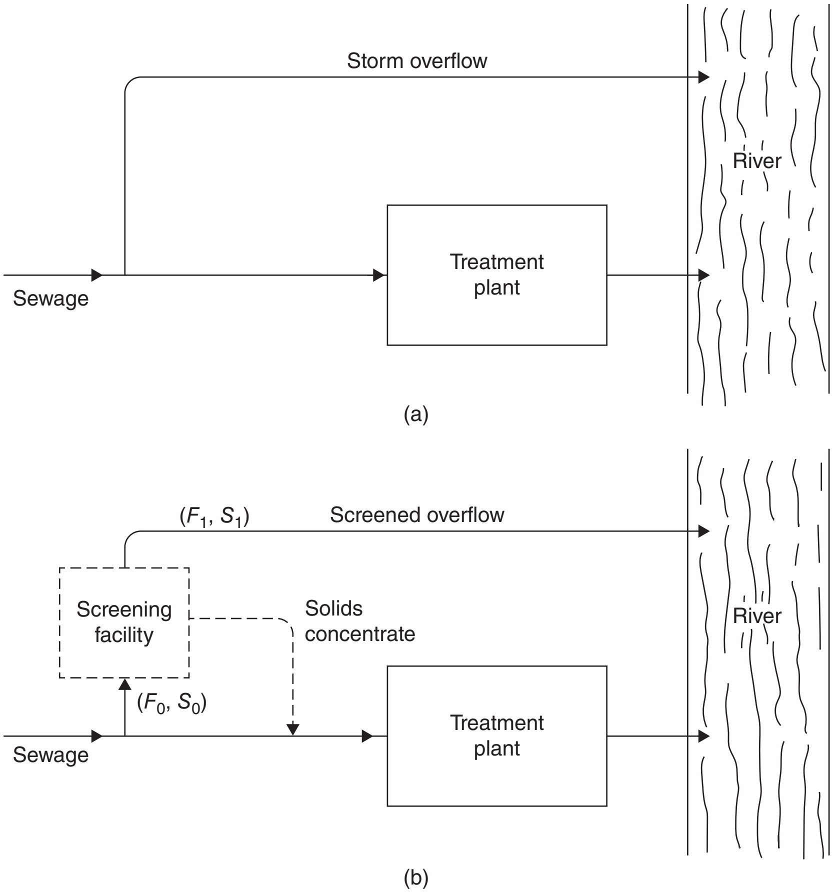


Figure 1.1.1 Operation of the sewage treatment plant: (a) standard mode of operation and (b) modified mode of operation, with screening facility, $F =$ flow; $S =$ settleable solids.

1.1.2 Investigation

The objective of the investigation was to determine good operating conditions.

1.1.3 Changing Criteria

What are good operating conditions? Initially, it was believed they were those resulting in the highest possible removal of solids. Referring to Figures 1.1.1b and 1.1.2a, settleable solids in the influent are denoted by S_0 and the settleable solids in the effluent by S_1. The *percent solids removed* by the screen is therefore $y = 100(S_0 - S_1)/S_0$. Thus, initially, it was believed that good operation meant achieving a high value for y. However, it became evident after the first set of experiments were made, that the *percentage of the flow retreated* (flow returned to treatment plant), which we denote by z, also had to be taken into account. Referring to Figures 1.1.1b and 1.1.2a, influent flow to the screens is denoted by F_0 and effluent flow from the screens to the river by F_1. Thus, $z = 100(F_0 - F_1)/F_0$.

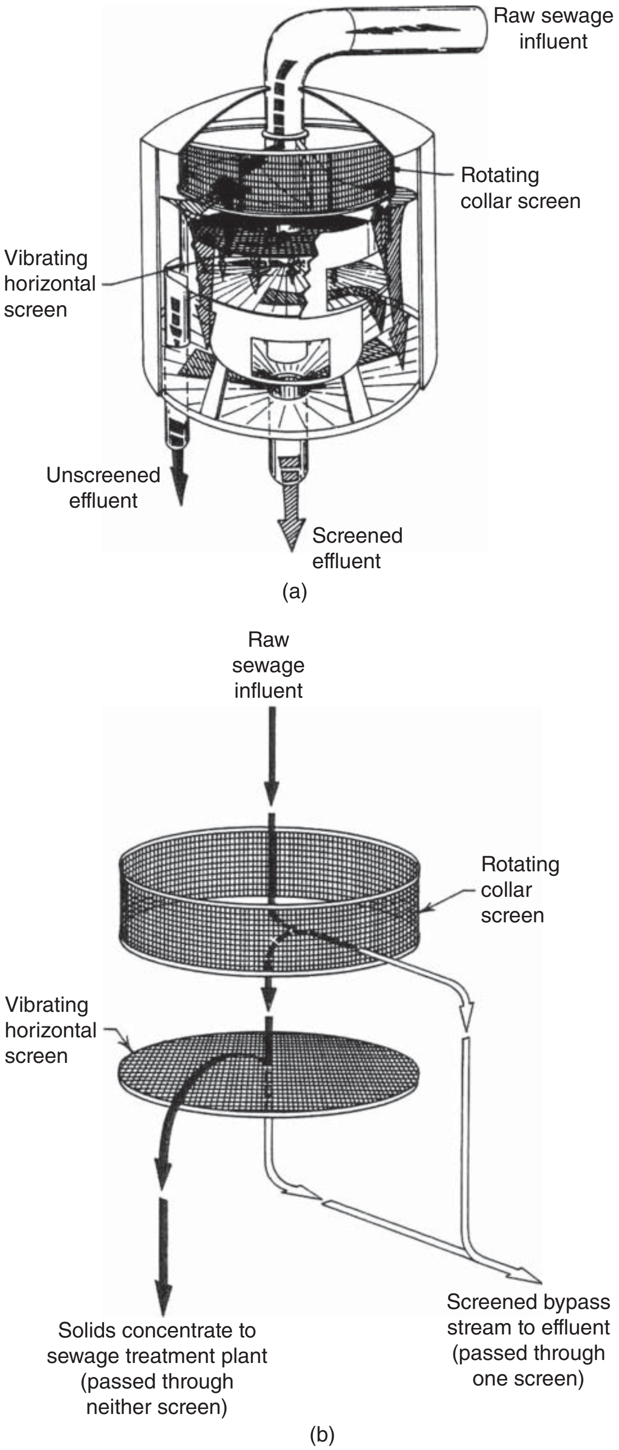

Raw sewage influent
Rotating collar screen
Vibrating horizontal screen
Unscreened effluent
Screened effluent
(a)
Raw sewage influent
Rotating collar screen
Vibrating horizontal screen
Solids concentrate to sewage treatment plant (passed through neither screen)
Screened bypass stream to effluent (passed through one screen)
(b)

Figure 1.1.2 Original version of the screening unit (a) detailed diagram and (b) simplified diagram.

1.1.4 A Summary of the Various Phases of the Investigation

Phase a

In this initial phase, an experiment was run in which the roles of three variables were studied: collar screen mesh size (fine, coarse), horizontal screen mesh size (fine, coarse), and flow rate (gpm). At this stage,

1. The experimenters were encouraged by the generally high values achieved for y.
2. Highest values for y were apparently achieved by using a horizontal screen with a coarse mesh and a collar screen with fine mesh.
3. Contrary to expectation, flow rate did not show up as an important variable affecting y.
4. Most important, the experiment was unexpectedly dominated by the z values, which measure the flow retreated. These were uniformly very low, with about 0.01% of the flow being returned to the treatment plant and 99.9% leaving the screen for discharge into the river. Although it was desirable that the retreated flow be small, the z values were embarrassingly low. As the experimenters remarked, "[T]he horizontal screen produced a solid concentrate . . . dry enough to shovel . . . . This represented a waste of effort of concentrating because the concentrated solids were intended to *flow* from the units."

Phase b

It was now clear (i) that z as well as y were important and (ii) that z was too low. It was conjectured that the matters might be improved by removing the horizontal screen altogether. Another experiment was therefore performed with no horizontal screen. The speed of rotation of the collar screen was introduced as a new variable.

Unfortunately, after only two runs of this experiment, this particular phase had to be terminated because of the excessive tearing of the cloth screens. From the scanty results obtained it appeared, however, that with no horizontal screen high solid removal could be achieved with a higher portion of the flow retreated. It was therefore decided to repeat these runs with screens made of stainless steel instead of cloth.

Phase c

A third experiment, using stainless steel collar screens of two mesh sizes, similar to that attempted in phase b, was performed with the same collar screen mesh size, collar screen speed (rpm), and flow rate (gpm) used before.

In this phase, with a stainless steel collar screen, high removal rates y were possible for eight sets of conditions for the factors just mentioned. However, these high y values were obtained with retreated flow z at undesirably high values (before, they had been too low). The objective was to get reasonably small values for z, but not so small as to make shoveling necessary; values between 5% and 20% were desirable. It was believed that by varying flow rate and speed of rotation of the collar screen, this objective could be achieved without sacrificing solid removal.

Phase d

Again, using a stainless steel collar screen, another experiment, with two factors, namely collar screen speed (rpm) and flow rate (gpm),set at two levels each, was run. This time, high values of solid removal were maintained, but unfortunately, flow retreated values were even higher than before.

Phase e

It was now conjectured that intermittent back washing could overcome the difficulties. This procedure was now introduced with influent flow rate and collar screen mesh varied.

The results of this experiment lead to a removal efficiency of $y = 89\%$ with a retreated flow of only $z = 8\%$. This was regarded as a satisfactory and practical solution, and the investigation was terminated at that point.

For detailed analysis of this experiment, the reader should refer to Box et al. (1978, p. 354). Of course, these types of experiments and their analyses are discussed in this text (see Chapter 18).

1.2 A SURVEY

The purpose of a sample survey is to make inferences about certain characteristics of a population from which samples are drawn. The inferences to be made for a population usually entails the estimation of population parameters, such as the population total, the mean, or the population proportion of a certain characteristic of interest. In any sample survey, a clear statement of its objective is very important. Without a clear statement about the objectives, it is very easy to miss pertinent information while planning the survey that can cause difficulties at the end of the study.

In any sample survey, only relevant information should be collected. Sometimes trying to collect too much information may become very confusing and consequently hinder the determination of the final goal. Moreover, collecting information in sample surveys costs money, so that the interested party must determine which and how much information should be obtained. For example, it is important to describe how much precision in the final results is desired. Too little information may prevent obtaining good estimates with desired precision, while too much information may not be needed and may unnecessarily cost too much money. One way to avoid such problems is to select an appropriate method of sampling the population. In other words, the sample survey needs to be appropriately designed. A brief discussion of such designs is given in Chapter 2. For more details on these designs, the reader may refer to Cochran (1977), Sukhatme and Sukhatme (1970), or Scheaffer et al. (2006).

1.3 AN OBSERVATIONAL STUDY

An observational study is one that does not involve any experimental studies. Consequently, observational studies do not control any variables. For example, a realtor wishes to appraise a house value. All the data used for this purpose are observational data. Many psychiatric studies involve observational data.

Frequently, in fitting a regression model (see Chapters 15 and 16), we use observational data. Similarly, in quality control (see Chapters 20 and 21), most of the data used in studying control charts for attributes are observational data. Note that control charts for attributes usually do not provide any cause-and-effect relationships. This is because observational data give us very limited information about cause-and-effect relationships.

As another example, many psychiatric studies involve observational data, and such data do not provide the cause of patient's psychiatric problems. An advantage of observational studies is that they are usually more cost-effective than experimental studies. The disadvantage of observational studies is that the data may not be as informative as experimental data.

1.4 A SET OF HISTORICAL DATA

Historical data are not collected by the experimenter. The data are made available to him/her.

Many fields of study such as the many branches of business studies, use historical data. A financial advisor for planning purposes uses sets of historical data. Many investment services provide financial data on a company-by-company basis.

1.5 A BRIEF DESCRIPTION OF WHAT IS COVERED IN THIS BOOK

Data collection is very important since it can greatly influence the final outcome of subsequent data analyses. After collection of the data, it is important to organize, summarize, present the preliminary outcomes, and interpret them. Various types of tables and graphs that summarize the data are presented in Chapter 2. Also in that chapter, we give some methods used to determine certain quantities, called *statistics*, which are used to summarize some of the key properties of the data.

The basic principles of probability are necessary to study various probability distributions. We present the basic principles of elementary probability theory in Chapter 3. Probability distributions are fundamental in the development of the various techniques of statistical inference. The concept of random variables is also discussed in Chapter 3.

Chapters 4 and 5 are devoted to some of the important discrete distributions, continuous distributions, and their moment-generating functions. In addition, we study in Chapter 5 some special distributions that are used in reliability theory.

In Chapter 6, we study joint distributions of two or more discrete and continuous random variables and their moment-generating functions. Included in Chapter 6 is the study of the bivariate normal distribution.

Chapter 7 is devoted to the probability distributions of some sample statistics, such as the sample mean, sample proportions, and sample variance. In this chapter, we also study a fundamental result of probability theory, known as the Central Limit Theorem. This theorem can be used to approximate the probability distribution of the sample mean when the sample size is large. In this chapter, we also study some sampling distributions of some sample statistics for the special case in which the population distribution is the so-called normal distribution. In addition, we present probability distributions of various "order statistics," such as the largest element in a sample, smallest element in a sample, and sample median.

Chapter 8 discusses the use of sample data for estimating the unknown population parameters of interest, such as the population mean, population variance, and population proportion. Chapter 8 also discusses the methods of estimating the difference of two population means, the difference of two population proportions, and the ratio of two population variances and standard deviations. Two types of estimators are included, namely point estimators and interval estimators (confidence intervals).

Chapter 9 deals with the important topic of statistical tests of hypotheses and discusses test procedures when concerned with the population means, population variance, and population proportion for one and two populations. Methods of testing hypotheses using the confidence intervals studied in Chapter 8 are also presented.

Chapter 10 gives an introduction to the theory of reliability. Methods of estimation and hypothesis testing using the exponential and Weibull distributions are presented.

In Chapter 11, we introduce the topic of data mining. It includes concepts of big data and starting steps in data mining. Classification, machine learning, and inference versus prediction are also discussed.

In Chapter 12, we introduce topic of cluster analysis. Clustering concepts and similarity measures are introduced. The hierarchical and nonhierarchical clustering techniques and model-based clustering methods are discussed in detail.

Chapter 13 is concerned with the chi-square goodness-of-fit test, which is used to test whether a set of sample data support the hypothesis that the sampled population follows some specified probability model. In addition, we apply the chi-square goodness-of-fit test for testing hypotheses of independence and homogeneity. These tests involve methods of comparing observed frequencies with those that are expected if a certain hypothesis is true.

Chapter 14 gives a brief look at tests known as "nonparametric tests," which are used when the assumption about the underlying distribution having some specified parametric form cannot be made.

Chapter 15 introduces an important topic of applied statistics: simple linear regression analysis. Linear regression analysis is frequently used by engineers, social scientists, health researchers, and biological scientists. This statistical technique explores the relation between two variables so that one variable can be predicted from the other. In this chapter, we discuss the least squares method for estimating the simple linear regression model, called the fitting of this regression model. Also, we discuss how to perform a residual analysis, which is used to check the adequacy of the regression model, and study certain transformations that are used when the model is not adequate.

Chapter 16 extends the results of Chapter 15 to multiple linear regressions. Similar to the simple linear regression model, multiple linear regression analysis is widely used. It provides statistical techniques that explore the relations among more than two variables, so that one variable can be predicted from the use of the other variables. In this chapter, we give a discussion of multiple linear regression, including the matrix approach. Finally, a brief discussion of logistic regression is given.

In Chapter 17, we introduce the design and analysis of experiments using one, two, or more factors. Designs for eliminating the effects of one or two nuisance variables along with a method of estimating one or more missing observations are given. We include two nonparametric tests, the Kruskal–Wallis and the Friedman test, for analyzing one-way and randomized complete block designs. Finally, models with fixed effects, mixed effects, and random effects are also discussed.

Chapter 18 introduces a special class of designs, the so-called 2^k factorial designs. These designs are widely used in various industrial and scientific applications. An extensive discussion of unreplicated 2^k factorial designs, blocking of 2^k factorial designs, confounding in the 2^k factorial designs, and Yates's algorithm for the 2^k factorial designs is also included. We also devote a section to fractional factorial designs, discussing one-half and one-quarter replications of 2^k factorial designs.

In Chapter 19, we introduce the topic of response surface methodology (RSM). First-order and second-order designs used in RSM are discussed. Methods of determining optimum or near optimum points using the "method of steepest ascent" and the analysis of a fitted second-order response surface are also presented.

Chapters 20 and 21 are devoted to control charts for variables and attributes used in phase I and phase II of a process. "Phase I" refers to the initial stage of a new process, and "phase II" refers to a matured process. Control charts are used to determine whether a process involving manufacturing or service is "under statistical control" on the basis of information contained in a sequence of small samples of items of interest. Due to lack of space, these two chapters are not included in the text but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

All the chapters are supported by three popular statistical software packages, MINITAB, R, and JMP. The MINITAB and R are fully integrated into the text of each chapter, whereas JMP is given in an independent section, which is not included in the text but is available for download from the book website: www.wiley.com/college/gupta/statistics2e. Frequently, we use the same examples for the discussion of JMP as are used in the discussion of MINITAB and R. For the use of each of these software packages, no prior knowledge is assumed, since we give each step, from entering the data to the final analysis of such data under investigation. Finally, a section of case studies is included in almost all the chapters.

Part I

Fundamentals of Probability and Statistics

Chapter 2

DESCRIBING DATA GRAPHICALLY AND NUMERICALLY

The focus of this chapter is a discussion of methods for describing sets of data.

Topics Covered

- Basic concepts of a population and various types of sampling designs
- Classification of the types of data
- Organizing and summarizing qualitative and quantitative data
- Describing qualitative and quantitative data graphically
- Determining measures of centrality and measures of dispersion for a set of raw data
- Determining measures of centrality and measures of dispersion for grouped data
- Determining measures of relative position
- Constructing a box whisker plot and its use in data analysis
- Determining measures of association
- Using statistical packages MINITAB, R, and JMP

Learning Outcomes

After studying this chapter, the reader will be able to do the following:

- Select an appropriate sampling design for data collection.
- Identify suitable variables in a problem and determine the level of measurement.
- Organize, summarize, present, and interpret the data.
- Identify the difference between a parameter and a statistic.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

- Calculate measures of the data such as mean, mode, median, variance, standard deviation, coefficient of variation, and measure of association and interpret them.
- Identify outliers if they are present in the data.
- Apply the statistical packages MINITAB, R, and JMP to analyze various sets of data.

2.1 GETTING STARTED WITH STATISTICS

2.1.1 What Is Statistics?

The term statistics is commonly used in two ways. On the one hand, we use the term statistics in day-to-day communication when we refer to the collection of numbers or facts. What follows are some examples of statistics:

1. In 2000, the salaries of CEOs from 10 selected companies ranged from $2 million to $5 million.
2. On average, the starting salary of engineers is 40% higher than that of technicians.
3. In 2007, over 45 million people in the United States did not have health insurance.
4. In 2008, the average tuition of private colleges soared to over $40,000.
5. In the United States, seniors spend a significant portion of their income on health care.
6. The R&D budget of the pharmaceutical division of a company is higher than the R&D budget of its biomedical division.
7. In December 2009, a total of 43 states reported rising jobless rates.

On the other hand, statistics is a scientific subject that provides the techniques of collecting, organizing, summarizing, analyzing, and interpreting the results as input to make appropriate decisions. In a broad sense, the subject of statistics can be divided into two parts: *descriptive statistics* and *inferential statistics.*

Descriptive statistics uses techniques to organize, summarize, analyze, and interpret the information contained in a data set to draw conclusions that do not go beyond the boundaries of the data set. Inferential statistics uses techniques that allow us to draw conclusions about a large body of data based on the information obtained by analyzing a small portion of these data. In this book, we study both descriptive statistics and inferential statistics. This chapter discusses the topics of descriptive statistics. Chapters 3 through Chapter 7 are devoted to building the necessary tools needed to study inferential statistics, and the rest of the chapters are mostly dedicated to inferential statistics.

2.1.2 Population and Sample in a Statistical Study

In a very broad sense, statistics may be defined as the science of collecting and analyzing data. The tradition of collecting data is centuries old. In European countries, numerous government agencies started keeping records on births, deaths, and marriages about four centuries ago. However, scientific methods of analyzing such data are not old. Most of the advanced techniques of analyzing data have in fact been developed only in the twentieth century, and routine use of these techniques became possible only after the invention of modern computers.

During the last four decades, the use of advanced statistical techniques has increased exponentially. The collection and analysis of various kinds of data has become essential in the fields of agriculture, pharmaceuticals, business, medicine, engineering, manufacturing, product distribution, and by government or nongovernment agencies. In a typical field, there is often need to collect quantitative information on all elements of interest, which is usually referred to as the *population*. The problem, however, with collecting all conceivable values of interest on all elements is that populations are usually so large that examining each element is not feasible. For instance, suppose that we are interested in determining the breaking strength of the filament in a type of electric bulb manufactured by a particular company. Clearly, in this case, examining each and every bulb means that we have to wait until each bulb dies. Thus, it is unreasonable to collect data on all the elements of interest. In other cases, as doing so may be either quite expensive, time-consuming, or both, we cannot examine all the elements. Thus, we always end up examining only a small portion of a population that is usually referred to as a *sample*. More formally, we may define population and sample as follows:

Definition 2.1.1 A *population* is a collection of all elements that possess a characteristic of interest.

Populations can be finite or infinite. A population where all the elements are easily countable may be considered as *finite*, and a population where all the elements are not easily countable as *infinite*. For example, a production batch of ball bearings may be considered a finite population, whereas all the ball bearings that may be produced from a certain manufacturing line are considered conceptually as being infinite.

Definition 2.1.2 A portion of a population selected for study is called a *sample*.

Definition 2.1.3 The *target population* is the population about which we want to make inferences based on the information contained in a sample.

Definition 2.1.4 The population from which a sample is being selected is called a *sampled population*.

The population from which a sample is being selected is called a *sampled population*, and the population being studied is called the *target population*. Usually, these two populations coincide, since every effort should be made to ensure that the sampled population is the same as the target population. However, whether for financial reasons, a time constraint, a part of the population not being easily accessible, the unexpected loss of a part of the population, and so forth, we may have situations where the sampled population is not equivalent to the whole target population. In such cases, conclusions made about the sampled population are not usually applicable to the target population.

In almost all statistical studies, the conclusions about a population are based on the information drawn from a sample. In order to obtain useful information about a population by studying a sample, it is important that the sample be a representative sample; that is, the sample should possess the characteristics of the population under investigation. For example, if we are interested in studying the family incomes in the United States, then our sample must consist of representative families that are very poor, poor, middle class, rich, and very rich. One way to achieve this goal is by taking a random sample.

Definition 2.1.5 A sample is called a *simple random sample* if each element of the population has the same chance of being included in the sample.

There are several techniques of selecting a random sample, but the concept that each element of the population has the same chance of being included in a sample forms the basis of all random sampling, namely *simple random sampling, systematic random sampling, stratified random sampling, and cluster random sampling.* These four different types of sampling schemes are usually referred to as *sample designs.*

Since collecting each data point costs time and money, it is important that in taking a sample, some balance be kept between the sample size and resources available. Too small a sample may not provide much useful information, but too large a sample may result in a waste of resources. Thus, it is very important that in any sampling procedure, an appropriate sampling design is selected. In this section, we will review, very briefly, the four sample designs mentioned previously.

Before taking any sample, we need to divide the *target population* into nonoverlapping units, usually known as *sampling units.* It is important to recognize that the sampling units in a given population may not always be the same. Sampling units are in fact determined by the sample design chosen. For example, in sampling voters in a metropolitan area, the sampling units might be individual voters, all voters in a family, all voters living in a town block, or all voters in a town. Similarly, in sampling parts from a manufacturing plant, the sampling units might be an individual part or a box containing several parts.

Definition 2.1.6 A list of all sampling units is called the *sampling frame.*

The most commonly used sample design is the *simple random sampling design*, which consists of selecting n (sample size) sampling units in such a way that each sampling unit has the same chance of being selected. If, however, the population is finite of size N, say, then the simple random sampling design may be defined as selecting n sampling units in such a way that each possible sample of size n has the same chance of being selected. The number of such samples of size n that may be formed from a finite population of size N is discussed in Section 3.4.3.

Example 2.1.1 (Simple random sampling) *Suppose that an engineer wants to take a sample of machine parts manufactured during a shift at a given plant. Since the parts from which the engineer wants to take the sample are manufactured during the same shift at the same plant, it is quite safe to assume that all parts are representative. Hence in this case, a* simple random sampling *design should be appropriate.*

The second sampling design is the *stratified random sampling design*, which may give improved results for the same amount of money spent for simple random sampling. However, a stratified random sampling design is appropriate when a population can be divided into various nonoverlapping groups called *strata*. The sampling units in each stratum are similar but differ from stratum to stratum. Each stratum is treated as a subpopulation, and a simple random sample is taken from each of these subpopulations or strata.

In the manufacturing world, this type of sampling situation arises quite often. For instance, in Example 2.1.1, if the sample is taken from a population of parts manufactured either in different plants or in different shifts, then stratified random sampling can be more appropriate than simple random sampling. In addition, there is the advantage of administrative convenience. For example, if the machine parts are manufactured in plants located in different parts of the country, then stratified random sampling can be beneficial. Often, each plant (stratum) has a sampling department that can conduct the random sampling within each plant. In order to obtain best results in this case, the sampling departments in all the plants need to communicate with one another before sampling in order to ensure that the same sampling norms are followed. Another example of stratified random sampling in manufacturing occurs when samples are taken of products that are produced in different batches; here, products produced in different batches constitute the different strata.

A third kind of sampling design is *systematic random sampling*. The systematic random sampling procedure is the easiest one. This sampling scheme is particularly useful in manufacturing processes, when the sampling is done from a continuously operating assembly line. Under this scheme, a first item is selected randomly and thereafter every mth ($m = N/n$) item manufactured is selected until we have a sample of the desired size (n). Systematic sampling is not only easy to employ but, under certain conditions, is also more precise than simple random sampling.

The fourth and last sampling design is *cluster random sampling*. In cluster sampling, each sampling unit is a group of smaller units. In the manufacturing environment, this sampling scheme is particularly useful since it is difficult to prepare a list of each part that constitutes a frame. On the other hand, it may be easier to prepare a list of boxes in which each box contains many parts. Thus, in this case, a cluster random sample is merely a simple random sample of these boxes. Another advantage of cluster sampling is that by selecting a simple random sample of only a few clusters, we can in fact have quite a large sample of smaller units. Such sampling is achieved at minimum cost, since both preparing the frame and taking the sample are much more economical. In preparing any frame, we must define precisely the characteristic of interest or variable, where a variable may be defined as follows:

Definition 2.1.7 A *variable* is a characteristic of interest that may take different values for different elements.

For example, an instructor is interested in finding the ages, heights, weights, GPA, gender, and family incomes of all the students in her engineering class. Thus, in this example, the variables (characteristics of interest) are ages, heights, weights, GPA, gender, and family incomes.

2.2 CLASSIFICATION OF VARIOUS TYPES OF DATA

In practice, it is common to collect a large amount of nonnumerical and/or numerical data on a daily basis. For example, we may collect data concerning customer satisfaction, comments of employees, or perceptions of suppliers. Or we may track the number of employees in various departments of a company or check weekly production volume in units produced and sales dollars per unit of time, and so on. All the data collected, however, cannot be treated the same way as there are differences in types of data. Accordingly, statistical data can normally be divided into two major categories:

- *Qualitative*
- *Quantitative*

Each of these categories can be further subdivided into two subcategories each. The two subcategories of qualitative data are *nominal* and *ordinal*, whereas the two subcategories of quantitative data are *interval* and *ratio*. We may summarize this classification of statistical data as in Figure 2.2.1.

The classification of data as nominal, ordinal, interval, and ratio is arranged in the order of the amount of information they can provide. Nominal data provide minimum information, whereas ratio data provide maximum information.

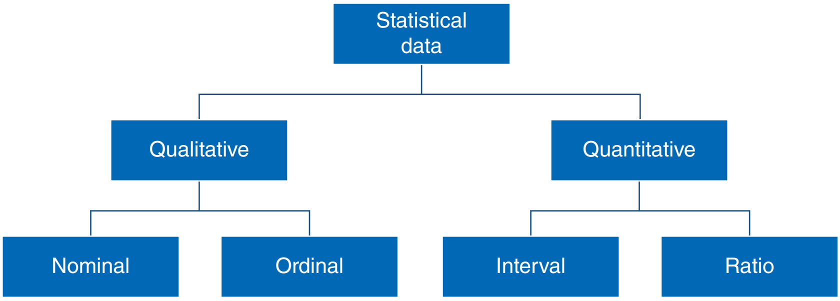


Figure 2.2.1 Classifications of statistical data.

2.2.1 Nominal Data

As previously mentioned, nominal data contain the smallest amount of information. Only symbols are used to label categories of a population. For example, production part numbers with a 2003 prefix are nominal data, wherein the 2003 prefix indicates only that the parts were produced in 2003 (in this case, the year 2003 serves as the category). No arithmetic operation, such as addition, subtraction, multiplication, or division, can be performed on numbers representing nominal data. As another example, jersey numbers of baseball, football, or soccer players are nominal. Thus, adding any two jersey numbers and comparing with another number makes no sense. Other examples of nominal data are ID numbers of workers, account numbers used by a financial institution, ZIP codes, telephone numbers, sex, or color.

2.2.2 Ordinal Data

Ordinal data are more informative than nominal data. When the ordering of categories becomes important, the data collected are called ordinal. Examples include companies ranked according to the quality of their product, companies ranked based on their annual revenues, the severity of burn types, or stage of cancer among cancer-afflicted patients. Again, no addition, subtraction, multiplication, or division can be used on ordinal-type data.

Other examples of ordinal data are represented by geographical regions, say designated as A, B, C, and D for shipping purposes, or preference of vendors who can be called upon for service, or skill ratings of certain workers of a company, or in electronics engineering, the color-coded resistors, which represent ascending order data.

2.2.3 Interval Data

Interval data are numerical data, more informative than nominal and ordinal data but less informative than ratio data. A typical example of interval data is temperature (in Celsius and Fahrenheit). Arithmetic operations of addition and subtraction are applicable, but multiplication and division are not applicable. For example, the temperature of three consecutive parts A, B, and C during a selected step in a manufacturing process are 20°F, 60°F, and 30°F, respectively. Then we can say the temperature difference between parts A and B is different from the difference between parts B and C. Also we can say that part B is warmer than part A and part C is warmer than part A, but cooler than part B. However, it is physically meaningless to say that part B is three times as warm as part A and twice as warm as part C. Moreover, in interval data, zero does not have the conventional meaning of nothingness; it is just an arbitrary point on the scale of measurement. For instance, 0°F and 0°C (=32°F) have different values, and they are in fact the arbitrary points on different scales of measurements. Other examples of interval data are year in which a part is produced, students' numeric grades on a test, and date of birth.

2.2.4 Ratio Data

Ratio data are also numerical data that have the potential to produce the most meaningful information of all data types. All arithmetic operations are applicable on this type of data. Numerous examples of this type of data exist, such as height, weight, length of rods, diameter of a ball bearing, RPM of a motor, number of employees in a company, hourly wages, and annual growth rate of a company. In ratio data, the number zero equates to nothingness. In other words, the number zero means absence of the characteristics of interest.

PRACTICE PROBLEMS FOR SECTIONS 2.1 AND 2.2

1. Describe briefly the difference between a sample and a population. Give an example of a population and a sample.
2. Describe the difference between descriptive statistics and inferential statistics.
3. A university professor is interested in knowing the average GPA of a graduating class. The professor decided to record the GPA of only those students who were

in his/her class during the last semester before graduation. Using this information (the data), the professor estimates the average GPA of the graduating class using the average of the GPAs he/she collected. Describe the following:

(a) Population of interest
(b) Sample collected by the professor
(c) The variable of interest

4. Describe whether each of the following scenarios would result in qualitative or quantitative data:

(a) Time needed to finish a project by a technician
(b) Number of days of stay in a hospital by a patient after bypass surgery
(c) Average number of cars passing through a toll booth each day
(d) Types of beverages served by a restaurant
(e) Size of a rod used in a project
(f) Condition of a home for sale (excellent, good, fair, bad)
(g) Heights of basketball players
(h) Dose of medication prescribed by a physician to his/her patients
(i) Recorded temperatures of a tourist place during the month of January
(j) Ages of persons waiting in a physician's office
(k) Speed of a vehicle crossing George Washington Bridge in New York
(l) Amount of interest reported in a tax return
(m) Sizes of cars available at a rental company (full, medium, compact, small)
(n) Manufacturers of cars parked in a parking lot

5. Referring to Problem 4, classify the data in each case as nominal, ordinal, interval, or ratio.

6. A consumer protection agency conducts opinion polls to determine the quality (excellent, good, fair, bad) of products imported from an Asian country. Suppose that the agency conducted a poll in which 1000 randomly selected individuals were contacted by telephone.

(a) What is the population of interest?
(b) What is the sample?
(c) Classify the variable of interest as nominal, ordinal, interval, or ratio.

2.3 FREQUENCY DISTRIBUTION TABLES FOR QUALITATIVE AND QUANTITATIVE DATA

In statistical applications, we often encounter large quantities of messy data. To gain insight into the nature of the data, we often organize and summarize the data by constructing a table called a *frequency distribution table*. In any statistical application (as noted in Section 2.2), we can have data that are either qualitative or quantitative. Qualitative and quantitative are sometimes referred to as categorical or numerical data, respectively. In this section, we discuss the construction of a *frequency distribution table* when the data are qualitative or quantitative.

2.3.1 Qualitative Data

A frequency distribution table for qualitative data consists of two or more categories along with the numbers of the data that belong to each category. The number of data belonging to any particular category is called the frequency or count of that category. We illustrate the construction of a frequency distribution table when the data are qualitative with the following example.

Example 2.3.1 (Industrial revenue) *Consider a random sample of 110 small to midsize companies located in the midwestern region of the United States, and classify them according to their annual revenues (in millions of dollars). Then construct a frequency distribution table for the data obtained by this classification.*

Solution: We classify the annual revenues into five categories as follows: Under 250, 250–under 500, 500–under 750, 750–under 1000, 1000 or more. Then the data collected can be represented as shown in Table 2.3.1, where we have used the labels $1, 2, \ldots, 5$ for the above categories.

Table 2.3.1 Annual revenues of 110 small to midsize companies located in mid-western region of the United States.

1	4	3	5	3	4	1	2	3	4	3	1	5	3	4	2	1	1	4	5	3	2	5	2	5	2	1	2	3
3	2	1	2	5	3	2	1	1	2	1	2	4	5	3	5	1	3	1	2	1	4	1	4	5	4	1	1	2
4	1	4	1	2	4	3	3	4	1	4	1	4	1	2	1	5	3	1	5	2	1	2	3	1	2	2	1	1
2	1	5	3	2	5	5	2	5	4	3	5	2	3	2	3	5	2	3	5	5	2	3	2	5	1	4		

After tallying the data, we find that of the 110 companies, 28 belong in the first category, 26 in the second category, 20 in the third category, 16 in the fourth category, and 20 in the last category. Thus, a frequency distribution table for the data in Table 2.3.1 is as shown in Table 2.3.2.

Table 2.3.2 Frequency distribution for the data in Table 2.3.1.

Categories	Tally	Frequency or count	Cumulative frequency	Percentage	Cumulative percentage
1	///// ///// ///// ///// ///// ///	28	28	25.45	25.45
2	///// ///// ///// ///// ///// /	26	54	23.64	49.09
3	///// ///// ///// /////	20	74	18.18	67.27
4	///// ///// ///// /	16	90	14.55	81.82
5	///// ///// ///// /////	20	110	18.18	100.00
Total		110		100.00	

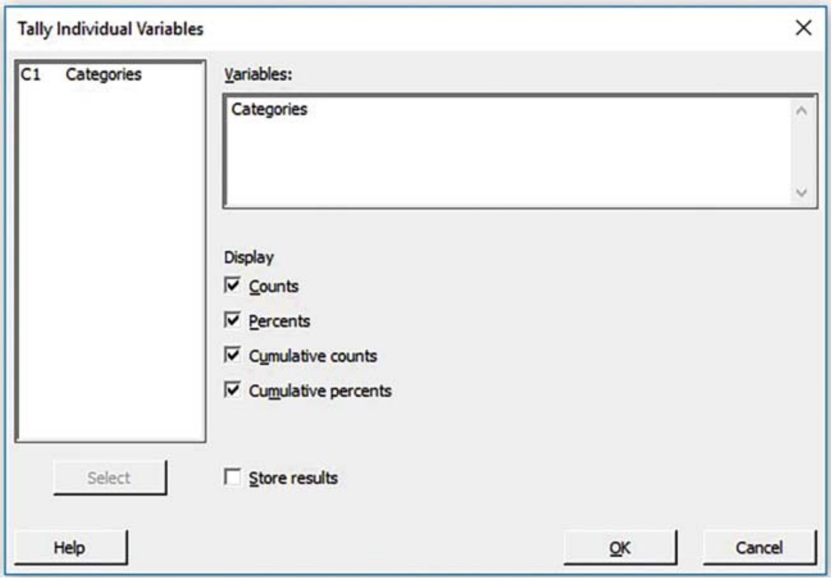
Interestingly, we can put technology to work on data in Table 2.3.1 to produce Table 2.3.2.

Example 2.3.2 (Industrial revenue) *Using MINITAB and R, construct a frequency distribution table for the data in Table 2.3.1.*

Solution:

MINITAB

1. Enter the data in column C1 of the Worksheet Window and name it Categories.
2. From the Menu bar, select **Stat** > **Tables** > **Tally Individual Variables**...



Tally Individual Variables

C1 Categories

Variables: Categories

Display

☑ Counts

☑ Percents

☑ Cumulative counts

☑ Cumulative percents

□ Store results

3. In this dialog box, enter C1 in the box under **Variables**.
4. Check all the boxes under Display and click **OK**.
5. The frequency distribution table as shown below appears in the Session window.

```
Categories  Count  Percent  CumCnt  CumPct
         1     28    25.45      28   25.45
         2     26    23.64      54   49.09
         3     20    18.18      74   67.27
         4     16    14.55      90   81.82
         5     20    18.18     110  100.00
        N =   110
```

This frequency distribution table may also be obtained by using R as follows:

USING R

R has built in 'table()' function that can be used to get the basic frequency distribution of categorical data. To get the cumulative frequencies, we can apply built in 'cumsum()' function to tabulated frequency data. Then using the 'cbind()' function we combine categories, frequencies, cumulative frequencies, and cumulative percentages to build the final distribution table. In addition, we can use the 'colnames()' function to name the columns of the final table as needed. The task can be completed running the following R code in R Console window.

```
#Assign given data to the variable data
data = c(4,3,5,3,4,1,2,3,4,3,1,5,3,4,2,1,1,4,5,3,2,5,2,5,2,1,2,3,3,2,
1,5,3,2,1,1,2,1,2,4,5,3,5,1,3,1,2,1,4,1,4,5,4,1,1,2,4,1,4,1,2,4,3,4,1,
4,1,4,1,2,1,5,3,1,5,2,1,2,3,1,2,2,1,1,2,1,5,3,2,5,5,2,5,3,5,2,3,2,3,5,
2,3,5,5,2,3,2,5,1,4)

#To get frequencies
data.freq = table(data)

#To combine necessary columns
freq.dist = cbind(data.freq, cumsum(data.freq), 100*cumsum(data.freq)/sum(data.freq))

#To name the table columns
colnames(freq.dist) = c('Frequency','Cum.Frequency','Cum Percentage')
freq.dist

#R output
```

```
     Frequency   Cum.Frequency   Cum Percentage
1        28.00           28.00            25.45
2        26.00           54.00            49.09
3        20.00           74.00            67.27
4        16.00           90.00            81.82
5        20.00          110.00           100.00
```

Note that sometimes a quantitative data set is such that it consists of only a few distinct observations that occur repeatedly. These kind of data are usually summarized in the same manner as the categorical data. The categories are represented by the distinct observations. We illustrate this scenario with the following example.

Example 2.3.3 (Hospital data) *The following data show the number of coronary artery bypass graft surgeries performed at a hospital in a 24-hour period for each of the last 50 days. Bypass surgeries are usually performed when a patient has multiple blockages or when the left main coronary artery is blocked. Construct a frequency distribution table for these data.*

1	2	1	5	4	2	3	1	5	4	3	4	6	2	3	3	2	2	3	5	2	5	3	4	3
1	3	2	2	4	2	6	1	2	6	6	1	4	5	4	1	4	2	1	2	5	2	2	4	3

Solution: In this example, the variable of interest is the number of bypass surgeries performed at a hospital in a period of 24 hours. Now, following the discussion in Example 2.3.1, we can see that the frequency distribution table for the data in this example is as shown in Table 2.3.3. Frequency distribution table defined by using a single numerical value is usually called a *single-valued frequency distribution* table.

Table 2.3.3 Frequency distribution table for the hospital data.

Categories	Tally	Frequency or count	Cumulative frequency	Percentage	Cumulative percentage
1	///// ///	8	8	16.00	16.00
2	///// ///// ////	14	22	28.00	44.00
3	///// ////	9	31	18.00	62.00
4	///// ////	9	40	18.00	80.00
5	///// /	6	46	12.00	92.00
6	////	4	50	8.00	100.00
Total		50		100.00	

2.3.2 Quantitative Data

So far, we have discussed frequency distribution tables for qualitative data and quantitative data that can be treated as qualitative data. In this section, we discuss frequency distribution tables for quantitative data.

Let $X_1, X_2, \ldots, X_n$ be a set of quantitative data values. To construct a frequency distribution table for this data set, we follow the steps given below.

Step 1. Find the range R of the data that is defined as

$$\text{Range} = R = \text{largest data point} - \text{smallest data point} \tag{2.3.1}$$

Step 2. Divide the data set into an appropriate number of *classes*. The classes are also sometimes called *categories*, *cells*, or *bins*. There are no hard and fast rules to determine the number of classes. As a rule, the number of classes, say m, should be somewhere between 5 and 20. However, *Sturges's* formula is often used, given by

$$\text{Number of classes} = m = 1 + 3.3 \log n \tag{2.3.2}$$

or

$$\text{Number of classes} = m = \sqrt{n} \tag{2.3.3}$$

where n is the total number of data points in a given data set and log denotes the log to base 10. The result often gives a good estimate for an appropriate number of intervals. Note that since m, the number of classes, should always be a whole number, the reader may have to round up or down the value of m obtained when using either equation (2.3.2) or (2.3.3).

Step 3. Determine the width of classes as follows:

$$\text{Class width} = R/m \tag{2.3.4}$$

The class width should always be a number that is easy to work with, preferably a whole number. Furthermore, this number should be obtained only by rounding up (never by rounding down) the value obtained when using equation (2.3.4).

Step 4. Finally, preparing the frequency distribution table is achieved by assigning each data point to an appropriate class. While assigning these data points to a class, one must be particularly careful to ensure that each data point be assigned to one, and only one, class and that the whole set of data is included in the table. Another important point is that the class at the lowest end of the scale must begin at a number that is less than or equal to the smallest data point and that the class at the highest end of the scale must end with a number that is greater than or equal to the largest data point in the data set.

Example 2.3.4 (Rod manufacturing) *The following data give the lengths (in millimeters) of 40 randomly selected rods manufactured by a company:*

145	140	120	110	135	150	130	132	137	115
142	115	130	124	139	133	118	127	144	143
131	120	117	129	148	130	121	136	133	147
147	128	142	147	152	122	120	145	126	151

Prepare a frequency distribution table for these data.

Solution: Following the steps described previously, we have the following:

1. Range $= R = 152 - 110 = 42$
2. Number of classes $= m = 1 + 3.3 \log 40 = 6.29 \approx 6$
3. Class width $= R/m = 42/6 = 7$

The six classes used to prepare the frequency distribution table are as follows: 110–under 117, 117–under 124, 124–under 131, 131–under 138, 138–under 145, 145–152.

Note that in the case of quantitative data, each class is defined by two numbers. The smaller of the two numbers is called the lower limit and the larger is called the upper limit. Also note that except for the last class, the upper limit does not belong to the class. For example, the data point 117 will be assigned to class two and not class one. Thus, no two classes have any common point, which ensures that each data point will belong to one and only one class. For simplification, we will use mathematical notation to denote the classes above as

$$[110\text{–}117),\ [117\text{–}124),\ [124\text{–}131),\ [131\text{–}138),\ [138\text{–}145),\ [145\text{–}152]$$

Here, the square bracket symbol "[" implies that the beginning point belongs to the class, and the parenthesis")" implies that the endpoint does not belong to the class. Then, the frequency distribution table for the data in this example is as shown in Table 2.3.4.

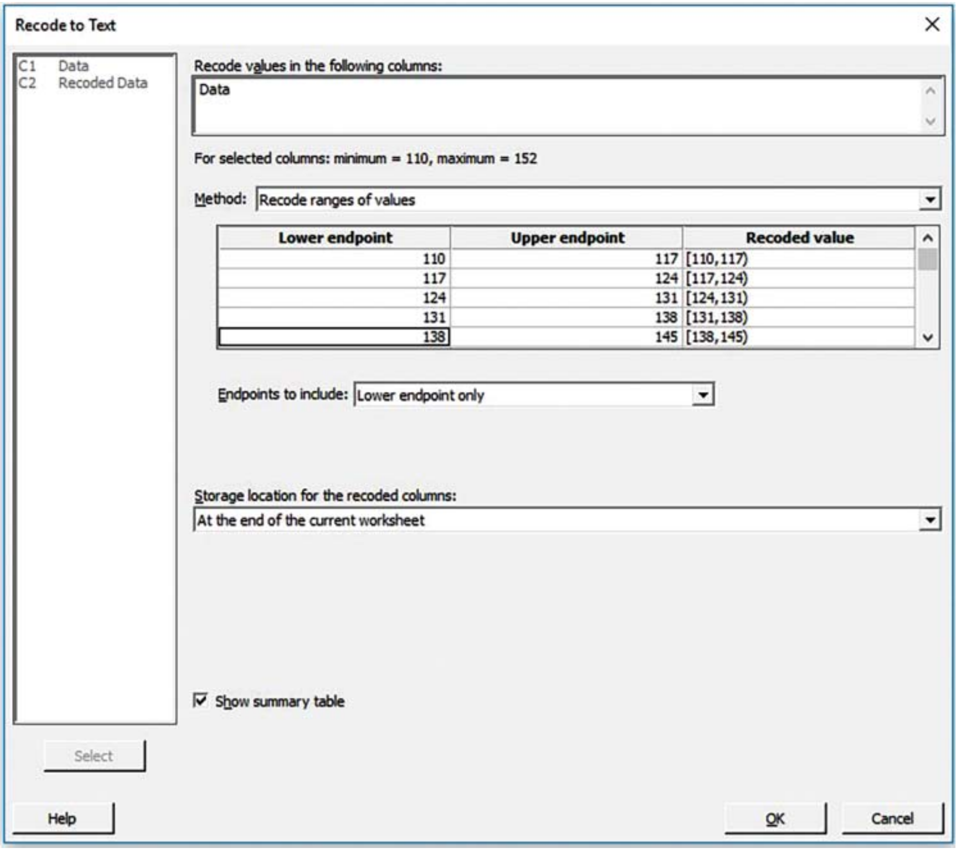
Table 2.3.4 Frequency table for the data on rod lengths.

Classes	Tally	Frequency or count	Relative frequency	Percentage	Cumulative frequency
[110 – –117)	///	3	3/40	7.5	3
[117 – –124)	///// //	7	7/40	17.5	10
[124 – –131)	///// ///	8	8/40	20.0	18
[131 – –138)	///// //	7	7/40	17.5	25
[138 – –145)	///// /	6	6/40	15.0	31
[145 – –152]	///// ////	9	9/40	22.5	40
Total		40	1	100	

The same frequency distribution table can be obtained by using MINITAB as follows:

MINITAB

1. Enter the data in column C1.
2. From the Menu bar select **Data** > **Recode** > **To Text**. This prompts the following dialog box to appear on the screen.



Recode to Text

C1 Data
C2 Recoded Data

Recode values in the following columns:
Data

For selected columns: minimum = 110, maximum = 152

Method: Recode ranges of values

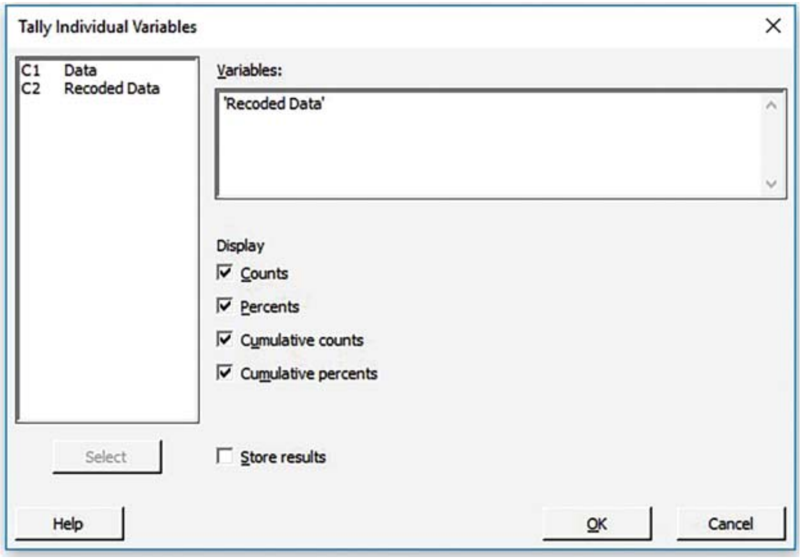
Lower endpoint	Upper endpoint	Recoded value
110	117	[110,117)
117	124	[117,124)
124	131	[124,131)
131	138	[131,138)
138	145	[138,145)

Endpoints to include: Lower endpoint only

Storage location for the recoded columns:
At the end of the current worksheet

☑ Show summary table

3. Enter C1 in the box under **Recode values in the following columns**.
4. Select Recode ranges of values from the pulldown menu next to **Method**.
5. Enter Lower and Upper endpoints as needed and make sure to change the final upper endpoint to 152.1. Type Recoded values in the interval format as previously shown.
6. Select **Lower endpoint only** from the pulldown menu next to **Endpoints to include**. Then, for the **Storage location for the recoded columns**, select **At the end of the current worksheet** from the pulldown menu.
7. Now from the Menu bar select **Stat** > **Tables** > **Tally Individual Variables**. This prompts the following dialog box to appear on the screen:



Tally Individual Variables

C1 Data
C2 Recoded Data

Variables: 'Recoded Data'

Display
☑ Counts
☑ Percents
☑ Cumulative counts
☑ Cumulative percents

□ Store results

8. In this dialog box, enter C2 Recoded Data in the box under **variables**.
9. Check all the boxes under **Display** and click **OK**. The frequency distribution table as shown below will appear in the Session window

```
Recoded Data   Count   Percent   CumCnt   CumPct
    [110,117)      3      7.50        3     7.50
    [117,124)      7     17.50       10    25.00
    [124,131)      8     20.00       18    45.00
    [131,138)      7     17.50       25    62.50
    [138,145)      6     15.00       31    77.50
    [145,152)      9     22.50       40   100.00
          N =     40
```

This frequency distribution table also can be obtained by using R as follows:

USING R

First, we define the required classes using the built in ‘seq()’ function. Then, we use the ‘cut()’ function to assign a corresponding class to each observation. As explained in

Example 2.3.2, we then use the ‘table()’ function on class variable and the ‘cusum()’ function on the frequency results. The task can be completed by running the following R code in the R Console window.

```
#Assign given data to the variable RodData
RodData = c(145,140,120,110,135,150,130,132,137,115,142,115,130,
124,139,133,118,127,144,143,131,120,117,129,148,130,121,136,133,
147,147,128,142,147,152,122,120,145,126,151)

#To define the intervals
breaks = seq(110, 152, by=7)

#To assign each observation to its interval
RodData.split = cut(RodData, breaks, right=FALSE)

#To obtain the frequency of data in each class
RodData.freq = table(RodData.split)

#To combine necessary columns
freq.dist = cbind(RodData.freq,100*RodData.freq/sum(RodData.freq),
cumsum(RodData.freq), 100*cumsum(RodData.freq)/sum(RodData.freq))

#To name the table columns
colnames(freq.dist) = c(‘Frequency’,‘Percentage’, ‘Cum.Frequency’,‘Cum.Percentage’)
freq.dist

#R output
```

```
            Frequency   Percentage   Cum.Frequency   Cum.Percentage
[110,117)        3.00         7.69            3.00             7.69
[117,124)        7.00        17.95           10.00            25.64
[124,131)        8.00        20.51           18.00            46.15
[131,138)        7.00        17.95           25.00            64.10
[138,145)        6.00        15.38           31.00            79.49
[145,152)        8.00        20.51           39.00           100.00
```

PRACTICE PROBLEMS FOR SECTION 2.3

1. The following data give the results of a customer sample survey for product satisfaction conducted by a manufacturing company. The numbers 1, 2, 3, 4, and 5 represent the satisfaction levels: very satisfied, fairly satisfied, neutral, fairly unsatisfied, and very unsatisfied, respectively.

1	1	3	3	4	2	4	3	1	5	1	2	2	4	1	1	2	4	4	2	5	4	2	2	1
3	4	4	5	5	2	3	2	3	2	3	2	4	4	3	1	5	1	5	4	1	1	1	5	2

(a) Prepare a frequency distribution table.
(b) Determine the percentages for all categories.
(c) What percentage of the customers in this sample survey was very satisfied or fairly satisfied?

2. An engineering school arranged a charity concert to raise funds for Iraq war veterans. The following data give the status of 40 randomly selected students who attended the concert. The numbers 1, 2, 3, and 4 represent the categories freshman, sophomore, junior, and senior, respectively.

2	3	4	1	1	2	1	1	3	2	1	3	3	4	4	4	1	1	3	2
3	3	3	3	3	4	3	3	3	4	2	3	2	2	3	1	4	2	2	3

(a) Prepare a frequency distribution table.
(b) Determine the percentages for all categories.
(c) What percentage of the students in this sample survey were juniors or seniors?

3. The following data give the responses of 36 senior citizens who were asked about the engine size of their car. The numbers 1, 2, 3, 4, and 5 represent the five categories 3.5, 3.2, 3.0, 2.2, and 1.8L, respectively.

5	4	5	5	1	2	1	1	1	3	5	1	5	4	2	1	3	1
3	2	3	4	1	2	2	1	5	5	3	1	5	2	1	2	2	5

(a) Prepare a frequency distribution table.
(b) Determine the percentages for all categories.
(c) What percentage of the senior citizens drive cars of category 1 or 3?

4. A manufacturing company of condenser retaining bolts for car engines implemented a quality control system. As part of this quality control system, a team of engineers decided to record the number of nonconforming bolts produced in each shift. The following data show the number of nonconforming bolts during the past 45 shifts.

25	30	26	26	25	16	21	22	21	27	15	24	19	20	24
16	30	28	24	23	15	15	21	28	18	15	21	27	26	28
17	19	24	26	27	17	27	19	22	27	16	25	16	30	18

Prepare a complete frequency distribution table, that is, a table having frequency, relative frequency, percentage, and cumulative frequency columns.

5. The following data give the number of graduate students admitted in all engineering programs of a prestigious university during the past 30 years (1976–2005).

148	167	171	177	175	165	134	177	168	142	126	166	130	122	157
138	163	129	143	145	141	162	147	141	164	137	149	146	132	157

Prepare a complete frequency distribution table, that is, a table having frequency, relative frequency, percentage, and cumulative frequency columns.

6. A temperature-sensing vacuum switch controls the vacuum that is applied to a vacuum motor operating a valve in the intake snorkel of the air cleaner. As the engine warms up, the temperature-sensing unit shuts off the vacuum applied to the motor, allowing the valve to close so that heated air shuts off and outside cooler air is drawn into the engine. The following data give the temperatures (coded) at which the sensing unit shuts off the vacuum:

105	101	120	116	108	112	118	119	107	100	107	120	113	113	101
102	102	100	101	100	118	106	114	100	104	101	107	113	110	100
109	108	100	104	110	113	118	100	119	120					

Prepare a complete frequency distribution table, that is, a table having frequency, relative frequency, percentage, and cumulative frequency columns.

2.4 GRAPHICAL DESCRIPTION OF QUALITATIVE AND QUANTITATIVE DATA

2.4.1 Dot Plot

A dot plot is one of the simplest graphs. To construct this graph, the value of each observation is plotted on a real line. It provides visual information about the distribution of a single variable. For illustration, we consider the following example.

Example 2.4.1 (Defective motors) *The following data give the number of defective motors received in 20 different shipments:*

8	12	10	16	10	25	21	15	17	5
26	21	29	8	6	21	10	17	15	13

Construct a dot plot for these data.

Solution: To construct a dot plot, draw a horizontal line with its scale beginning with a number less than the smallest observation and ending at a number greater than the largest

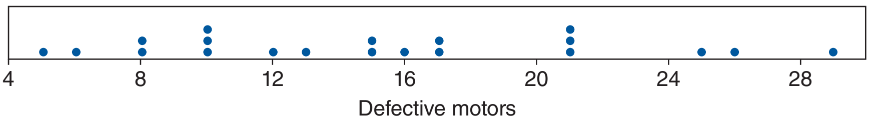


Figure 2.4.1 Dot plot for the data on defective motors received in 20 different shipments.

observation. Then, construct the dot plot by plotting the given data points as shown in Figure 2.4.1.

Dot plots are more useful when the sample size is small. A dot plot gives us, for example, information about how the data are scattered and where most of the observations are concentrated. For instance, in this example, we see that the minimum number of defective motors and the maximum number of defective motors received in any shipment were 5 and 29, respectively. Also, we can see that 75% of the time, the number of defective motors was between 8 and 21 (inclusive) for these shipments, and so on.

2.4.2 Pie Chart

Pie charts are commonly used to describe qualitative data from one population. It is constructed by dividing a circle into various slices that represent different categories of a population. As examples: allocation of the federal budget by sector, revenues of a large manufacturing company by region or by plant, technicians in a large corporation who are classified according to their basic qualification: high-school diploma, an associate degree, an undergraduate degree, a graduate degree, and so on. The pie chart helps us better understand at a glance the composition of the population with respect to the characteristic of interest.

To construct a pie chart, divide a circle into slices such that each slice representing a category is proportional to the size of that category. Since the total angle of a circle is 360°, the angle of a slice corresponding to a given category is determined as follows:

$$\text{Angle of a slice} = (\text{Relative frequency of the given category}) \times 360 \tag{2.4.1}$$

We illustrate the construction of a pie chart with the following example:

Example 2.4.2 (Manufacturing defect types) *In a manufacturing operation, we are interested in understanding defect rates as a function of various process steps. The inspection points (categories) in the process are initial cutoff, turning, drilling, and assembly. The frequency distribution table for these data is shown in Table 2.4.1. Construct a pie chart for these data.*

Table 2.4.1 Understanding defect rates as a function of various process steps.

Process steps	Frequency	Relative frequency	Angle size
Initial cutoff	86	86/361 = 23.8%	85.76
Turning	182	182/361 = 50.4%	181.50
Drilling	83	83/361 = 23.0%	82.77
Assembly	10	10/361 = 2.8%	9.97
Total	361	100%	360.00

Solution: The pie chart for these data is constructed by dividing the circle into four slices. The angle of each slice is given in the last column of Table 2.4.1. Then, the pie chart for the data of Table 2.4.1 is as shown in the MINITAB printout in Figure 2.4.2. Clearly, the pie chart gives us a better understanding at a glance about the rate of defects occurring at different steps of the process.

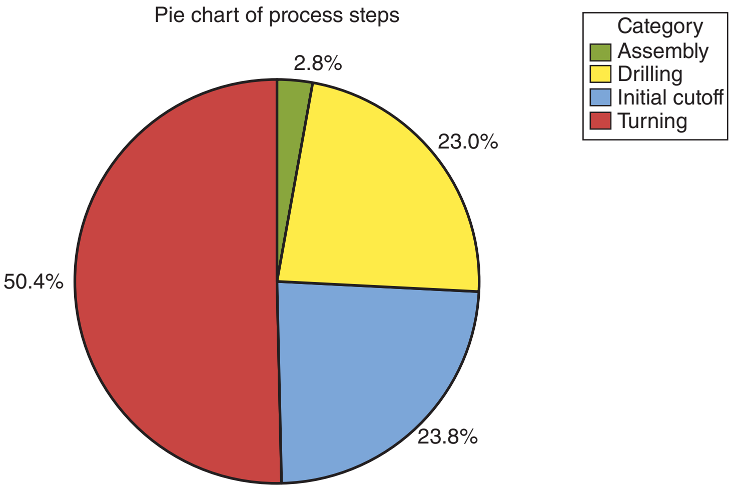
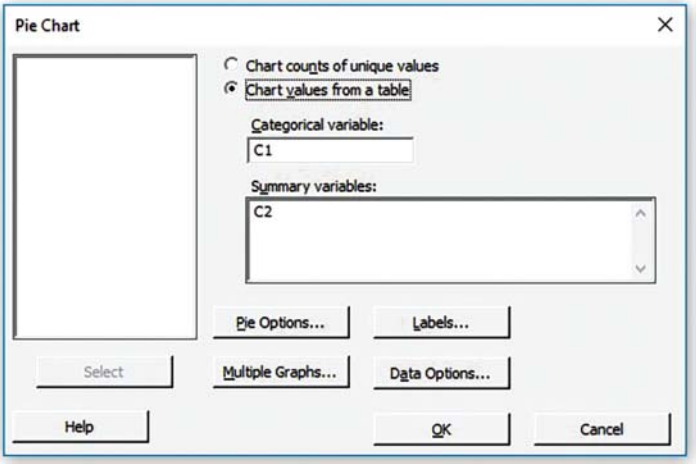


Figure 2.4.2 Pie chart for the data in Table 2.4.1 using MINITAB.

MINITAB

Using MINITAB, the pie chart is constructed by taking the following steps:

1. Enter the category in column C1.
2. Enter frequencies of the categories in column C2.
3. From the Menu bar, select **Graph** > **Pie Chart**. Then, check the circle next to **Chart values from a table** on the pie chart dialog box that appears on the screen.



Categorical variable: C1

Summary variables: C2

4. Enter C1 under **Categorical** values and C2 under **Summary variables**.
5. Note that if we have the raw data without having the frequencies for different categories, then check the circle next to **Chart counts of unique values**. In that case, the preceding dialog box would not contain a box for Summary variables.
6. Click **Pie Options** and in the new dialog box that appears select any option you like and click **OK**. Click **Lables** and in the new dialog box that appears select the

Slice Labels from the box menu and select **Percent** option and click **OK**. The pie chart will appear as shown in Figure 2.4.2.

USING R

We can use the built in ‘pie()’ function in R to generate pie charts. If a pie chart with percentages desired, then the percentages of the categories should be calculated manually. Then, these percentages should be used to label the categories. The task can be completed by running the following R code in the R Console window.

```
Freq = c(86, 182, 83, 10)

#To label categories
Process = c(‘Initial cutoff’, ‘Turning’, ‘Drilling’, ‘Assembly’)

#To calculate percentages
Percents = round(Freq/sum(Freq)*100,1)
label = paste(Percents, ‘%’, sep=‘ ’) # add % to labels

#Pie Chart with percentages
pie(Freq, labels = label, col=c(2,3,4,5), main=‘Pie Chart of Process Steps’)

#To add a legend. Note: “pch” specifies various point shapes.
legend(‘topleft’, Process, col=c(2,3,4,5), pch=15)
```

2.4.3 Bar Chart

Bar charts are commonly used to describe qualitative data classified into various categories based on sector, region, different time periods, or other such factors. Different sectors, different regions, or different time periods are then labeled as specific categories. A bar chart is constructed by creating categories that are represented by labeling each category and which are represented by intervals of equal length on a horizontal axis. The count or frequency within the corresponding category is represented by a bar of height proportional to the frequency. We illustrate the construction of a bar chart in the examples that follow.

Example 2.4.3 (Companies’ revenue) *The following data give the annual revenues (in millions of dollars) of five companies A, B, C, D, and E for the year 2011:*

78, 92, 95, 94, 102

Construct a bar chart for these data.

Solution: Following the previous discussion, we construct the bar chart as shown in Figure 2.4.3.

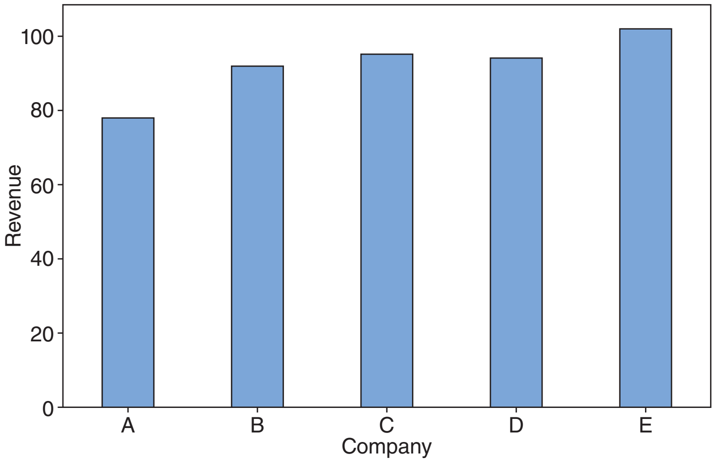


Figure 2.4.3 Bar chart for annual revenues of five companies for the year 2011.

Example 2.4.4 (Auto part defect types) *A company that manufactures auto parts is interested in studying the types of defects in parts produced at a particular plant. The following data shows the types of defects that occurred over a certain period:*

2	1	3	1	2	1	5	4	3	1	2	3	4	3	1	5	2	3	1	2	3	5	4	3	1
5	1	4	2	3	2	1	2	5	4	2	4	2	5	1	2	1	2	1	5	2	1	3	1	4

Construct a bar chart for the types of defects found in the auto parts.

Solution: In order to construct a bar chart for the data in this example, we first need to prepare a frequency distribution table. The data in this example are the defect types, namely 1, 2, 3, 4, and 5. The frequency distribution table is shown in Table 2.4.2. Note that the frequency distribution table also includes a column of cumulative frequency.

Now, to construct the bar chart, we label the intervals of equal length on the horizontal line with the category types of defects and then indicate the frequency of observations associated with each defect by a bar of height proportional to the corresponding frequency.

Table 2.4.2 Frequency distribution table for the data in Example 2.4.4.

Categories	Tally	Frequency or count	Relative frequency	Cumulative frequency
1	///// ///// ////	14	14/50	14
2	///// ///// ///	13	13/50	27
3	///// ////	9	9/50	36
4	///// //	7	7/50	43
5	///// //	7	7/50	50
Total		50	1.00	

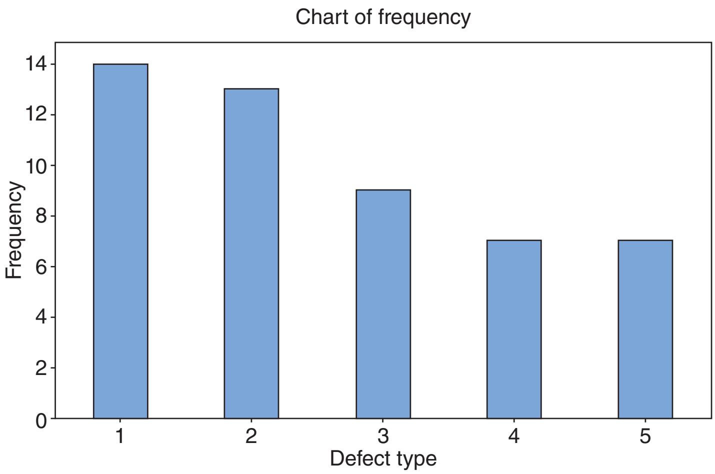


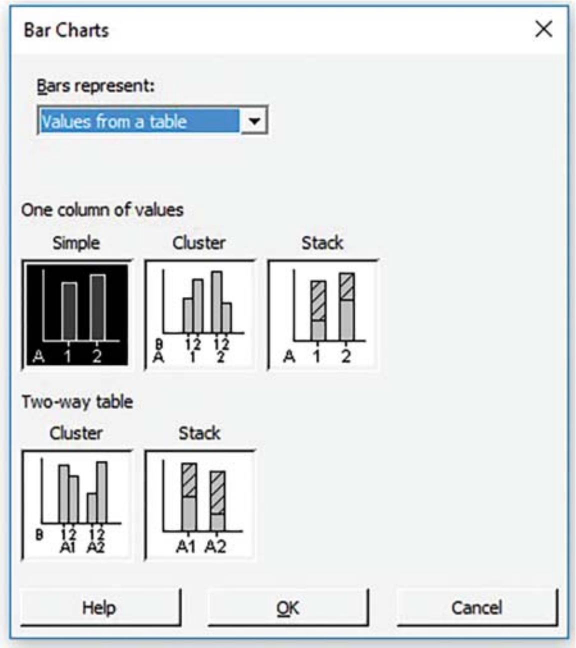
Figure 2.4.4 Bar graph for the data in Example 2.4.4.

Thus, the desired bar graph, as given in Figure 2.4.4, shows that the defects of type 1 occur the most frequently, type 2 occur the second most frequently, and so on.

MINITAB

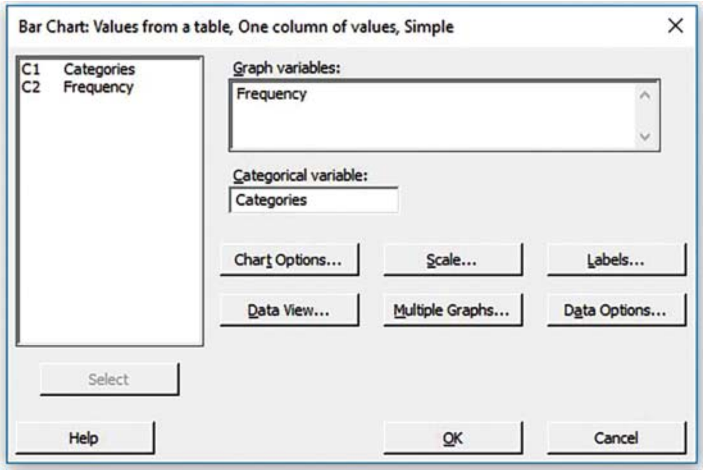
Using MINITAB, the bar chart is constructed by taking the following steps.

1. Enter the category in column C1.
2. Enter frequencies of the categories in C2.
3. From the Menu bar select **Graph** > **Bar Chart**. This prompts the following dialog box to appear on the screen:


Bar Charts

Bars represent: Values from a table

4. Select one of the three options under **Bars represent**, that is, **Counts of unique values, A function of variables**, or **Values from a table**, depending upon whether the data are sample values, functions of sample values such as means of various samples, or categories and their frequencies.
5. Select one of the three possible bar charts that suits your problem. If we are dealing with only one sample from a single population, then select **Simple** and click **OK**. This prompts another dialog box, as shown below, to appear on the screen:



Bar Chart: Values from a table, One column of values, Simple

C1 Categories
C2 Frequency

Graph variables: Frequency

Categorical variable: Categories

6. Enter C2 in the box under **Graph Variables**.
7. Enter C1 in the box under **Categorical values**.
8. There are several other options such as **Chart Option, scale**; click them and use them as needed. Otherwise click **OK**. The bar chart will appear identical to the one shown in Figure 2.4.4.

USING R

We can use built in ‘barplot()’ function in R to generate bar charts. First, we obtain the frequency table via the ‘table()’ function. The resulting tabulated categories and their frequencies are then inputted into the ‘barplot()’ function as shown in the following R code.

```
DefectTypes = c(2,1,3,1,2,1,5,4,3,1,2,3,4,3,1,5,2,3,1,2,3,5,4,3,
1,5,1,4,2,3,2,1,2,5,4,2,4,2,5,1,2,1,2,1,5,2,1,3,1,4)

#To obtain the frequencies
counts = table(DefectTypes)

#To obtain the bar chart
barplot(counts, xlab=‘Defect type’, ylab=‘Frequency’)
```

2.4.4 Histograms

Histograms are extremely powerful graphs that are used to describe quantitative data graphically. Since the shape of a histogram is determined by the frequency distribution table of the given set of data, the first step in constructing a histogram is to create a frequency distribution table. This means that a histogram is not uniquely defined until the classes or bins are defined for a given set of data. However, a carefully constructed histogram can be very informative.

For instance, a histogram provides information about the patterns, location/center, and dispersion of the data. This information is not usually apparent from raw data. We may define a histogram as follows:

Definition 2.4.1 A *histogram* is a graphical tool consisting of bars placed side by side on a set of intervals (classes, bins, or cells) of equal width. The bars represent the frequency or relative frequency of classes. The height of each bar is proportional to the frequency or relative frequency of the corresponding class.

To construct a histogram, we take the following steps:

Step 1. Prepare a frequency distribution table for the given data.
Step 2. Use the frequency distribution table prepared in Step 1 to construct the histogram. From here, the steps involved in constructing a histogram are exactly the same as those to construct a bar chart, except that in a histogram, there is no gap between the intervals marked on the horizontal axis (the x-axis).

A histogram is called a *frequency histogram* or a *relative frequency histogram* depending on whether the scale on the vertical axis (the y-axis) represents the frequencies or the relative frequencies. In both types of histograms, the widths of the rectangles are equal to the class width. The two types of histograms are in fact *identical* except that the scales used on the y-axes are different. This point becomes quite clear in the following example:

Example 2.4.5 (Survival times) *The following data give the survival times (in hours) of 50 parts involved in a field test under extraneous operating conditions.*

60	100	130	100	115	30	60	145	75	80	89	57	64	92	87	110	180
195	175	179	159	155	146	157	167	174	87	67	73	109	123	135	129	141
154	166	179	37	49	68	74	89	87	109	119	125	56	39	49	190	

Construct a frequency distribution table for this data. Then, construct frequency and relative frequency histograms for these data.

Solution:

Step 1. Find the range of the data:

$$R = 195 - 30 = 165$$

Then, determine the number of classes (see for example the Sturges' formula $m = 1 + 3.3 \log n$, in (2.3.2))

$$m = 1 + 3.3 \log 50 = 6.61 \approx 7$$

Last, compute the class width:

$$\text{Class width} = R/m = 165/7 = 23.57 \approx 24$$

As we noted earlier, the class width number is always rounded up to another convenient number that is easy to work with. If the number calculated using (2.3.4) is rounded down, then some of the observations will be left out as they will not belong to any class. Consequently, the total frequency will be less than the total count of the data. The frequency distribution table for the data in this example is shown in Table 2.4.3.

Step 2. Having completed the frequency distribution table, construct the histograms. To construct the frequency histogram, first mark the classes on the x-axis and the frequencies on the y-axis. Remember that when marking the classes and identifying the bins on the x-axis, there must be no gap between them. Then, on each class marked on the x-axis, place a rectangle, where the height of each rectangle is proportional to the frequency of the corresponding class. The frequency histogram for the data with the frequency distribution given in Table 2.4.3 is shown in Figure 2.4.5. To construct the relative frequency histogram, the scale is changed on the y-axis (see Figure 2.4.5) so that instead of plotting the frequencies, we plot relative frequencies. The resulting graph for this example, shown in Figure 2.4.6, is called the relative frequency histogram for the data with relative frequency distribution given in Table 2.4.3.

Table 2.4.3 Frequency distribution table for the survival time of parts.

Class	Tally	Frequency or count	Relative frequency	Cumulative frequency
[30 – –54)	/////	5	5/50	5
[54 – –78)	///// /////	10	10/50	15
[78 – –102)	///// ////	9	9/50	24
[102 – –126)	///// //	7	7/50	31
[126 – –150)	///// /	6	6/50	37
[150 – –174)	///// /	6	6/50	43
[174 – –198]	///// //	7	7/50	50
Total		50	1	

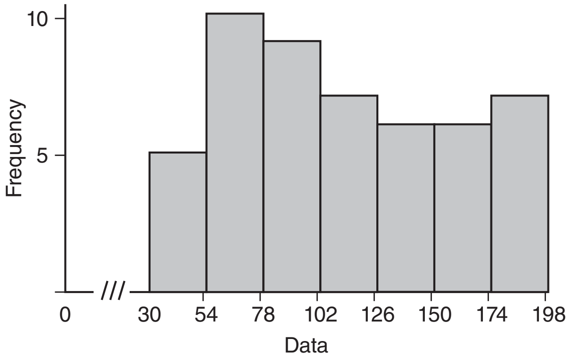


Figure 2.4.5 Frequency histogram for survival time of parts under extraneous operating conditions.

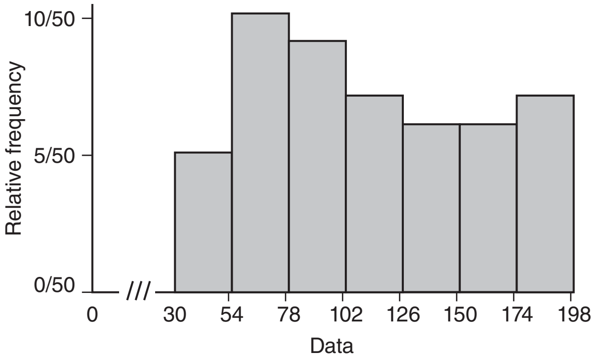


Figure 2.4.6 Relative frequency histogram for survival time of parts under extraneous operating conditions.

Another graph that becomes the basis of probability distributions, which we will study in later chapters, is called the *frequency polygon* or *relative frequency polygon* depending on which histogram is used to construct this graph. To construct the frequency or relative frequency polygon, first mark the midpoints on the top ends of the rectangles of the corresponding histogram and then simply join these midpoints. Note that classes with zero frequencies at the lower as well as at the upper end of the histogram are included so that we can connect the polygon with the x-axis. The lines obtained by joining the midpoints are called the *frequency* or *relative frequency polygons*, as the case may be. The frequency polygon for the data in Example 2.4.5 is shown in Figure 2.4.7. As the frequency and the relative frequency histograms are identical in shape, the frequency and relative frequency polygons are also identical, except for the labeling of the y-axis.

Quite often a data set consists of a large number of observations that result in a large number of classes of very small widths. In such cases, frequency polygons or relative frequency polygons become smooth curves. Figure 2.4.8 shows one such smooth curve. Such smooth curves, usually called frequency distribution curves, represent the probability distributions of continuous random variables that we study in Chapter 5. Thus, the histograms eventually become the basis for information about the probability distributions from which the sample was obtained.

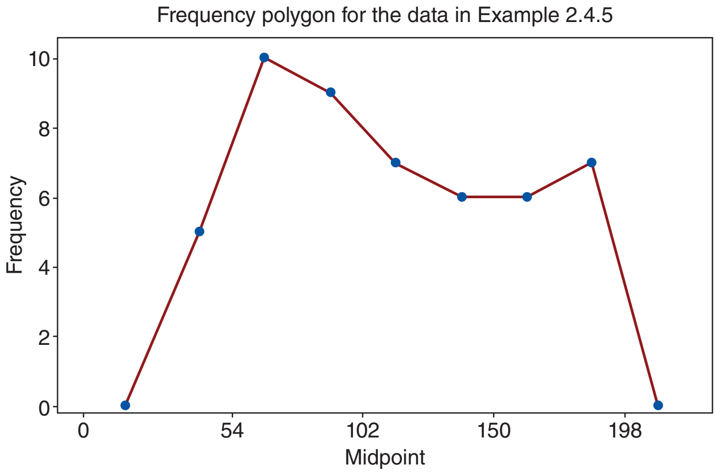


Figure 2.4.7 Frequency polygon for survival time of parts under extraneous operating conditions.

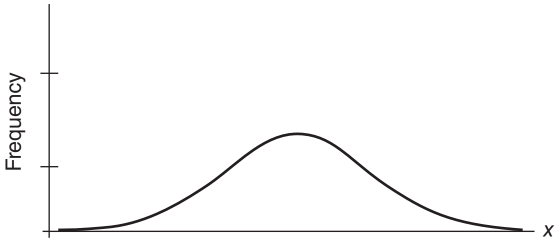


Figure 2.4.8 Typical frequency distribution curve.

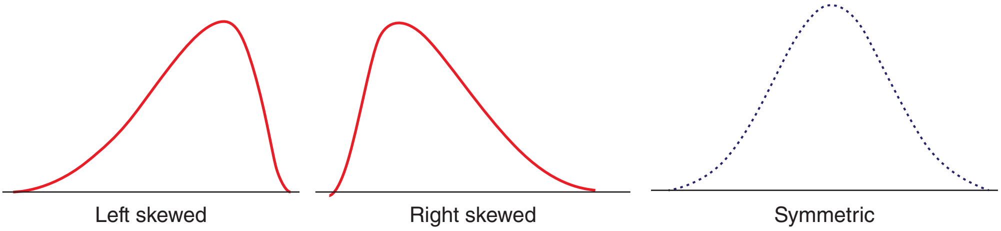


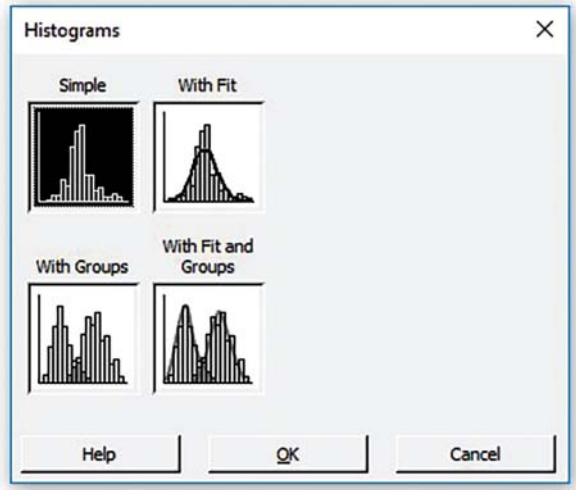
Figure 2.4.9 Three typical types of frequency distribution curves.

The shape of the frequency distribution curve of a data set depends on the shape of its histogram and choice of class or bin size. The shape of a frequency distribution curve can in fact be of any type, but in general, we encounter the three typical types of frequency distribution curves shown in Figure 2.4.9.

We now turn to outlining the various steps needed when using MINITAB and R.

MINITAB

1. Enter the data in column C1.
2. From the Menu bar, select **Graph** > **Histogram**. This prompts the following dialog box to appear on the screen.



3. From this dialog box, select an appropriate histogram and click **OK**. This will prompt another dialog box to appear.
4. In this dialog box, enter C1 in the box under the **Graph variables** and click **OK**. Then, a histogram graph will appear in the Session window.
5. After creating the histogram, if you want to customize the number of classes (cells or bins), click twice on any bar of the histogram. This prompts another dialog box **Edit Bars** to appear. In the new dialog box, select **Binning**. This allows the user to select the desired number of classes, their midpoints or cutpoints.

   To create a cumulative frequency histogram, take all the steps as previously described. Follow this in the dialog box at **Histogram-Simple**, and select **Scale** > **Y-Scale Type**. Then, check a circle next to **Frequency** and a square next to **Accumulate values across bins**. Click **OK**. A customized *Cumulative Frequency Histogram* using MINITAB is as obtained as shown in Figure 2.4.10. *Note*: To get the exact sample cumulative distribution, we used the manual cutpoints shown in the first column of Table 2.4.3 when Binning.
6. To obtain the frequency polygon in the dialog box **Histogram-Simple**, select **Data view** > **Data Display**, remove the check mark from **Bars**, and placing a check mark on **Symbols**. Under the **Smoother** tab, select **Lowess** for smoother and change **Degree of smoothing** to be 0 and **Number of steps** to be 1. Then, click **OK** twice. At this juncture, the polygon needs be modified to get the necessary cutpoints. We produced by right clicking on the X-axis, and selecting the **edit X scale**. Under the **Binning** tab for **Interval Type**, select **Cutpoint**, and under the **Interval Definition**, select **Midpoint/Cutpoint positions**. Now type manually calculated interval cutpoints. Note that one extra lower and one extra upper cutpoint should be included so that we can connect the polygon with the x-axis as shown in Figure 2.4.7.

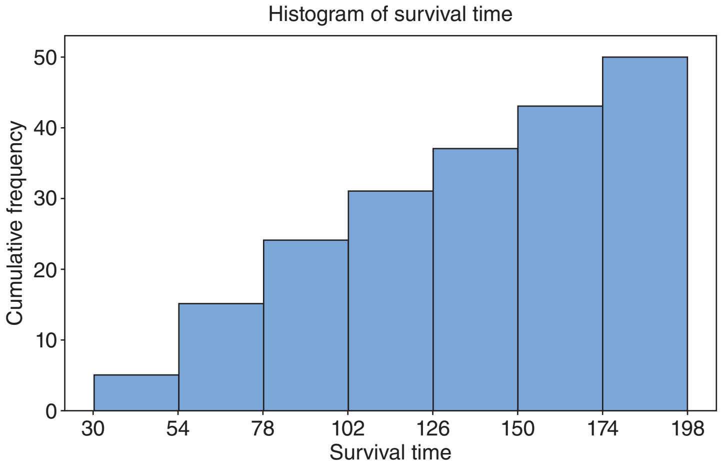
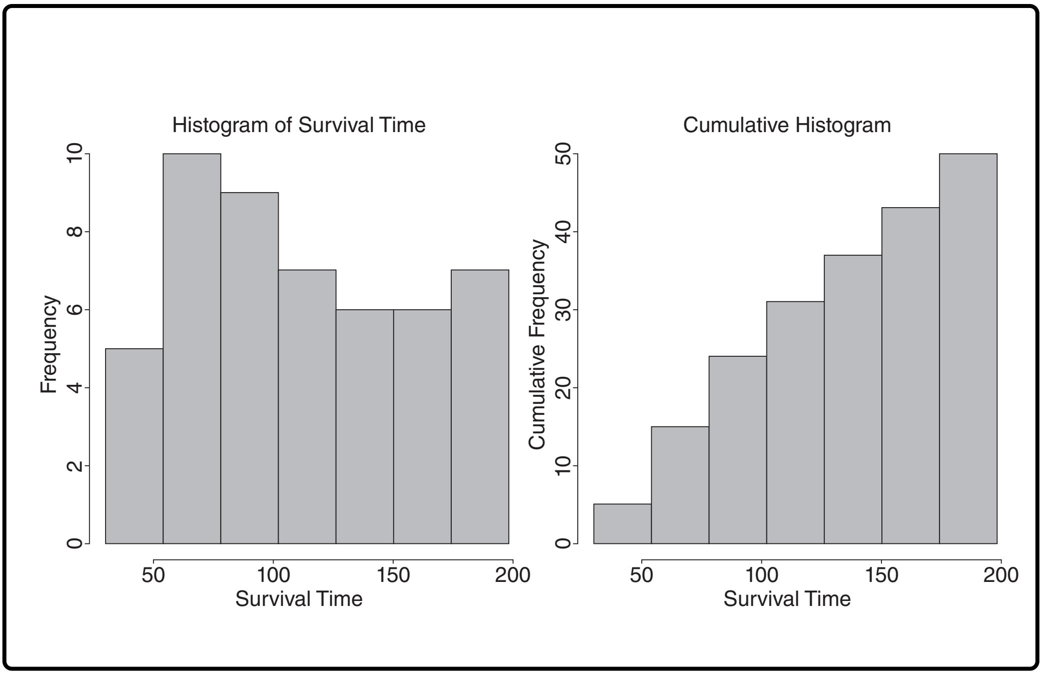


Figure 2.4.10 The cumulative frequency histogram for the data in Example 2.4.5.

USING R

We can use the built in ‘hist()’ function in R to generate histograms. Extra arguments such as ‘breaks’, ‘main’, ‘xlab’, ‘ylab’, ‘col’ can be used to define the break points, graph heading, x-axis label, y-axis label, and filling color, respectively. The specific argument ‘right = FALSE’ should be used to specify that the upper limit does not belong to the class. To obtain the cumulative histogram, we apply the ‘cumsum()’ function to frequencies obtained from the histogram function. The task can be completed by running the following R code in the R Console window.

```
SurvTime = c(60,100,130,100,115,30,60,145,75,80,89,57,64,92,87,110,
180,195,175,179,159,155, 146,157,167,174,87,67,73,109,123,135,129,
141,154,166,179,37,49,68,74,89,87,109,119,125,56,39,49,190)


#To plot the histogram
hist(SurvTime, breaks=seq(30,198, by=24), main=‘Histogram of Survival Time’,
xlab=‘Survival Time’, ylab=‘Frequency’, col=‘grey’, right = FALSE)

#To obtain the cumulative histogram, we replace cell
frequencies by their cumulative frequencies
h = hist(SurvTime, breaks=seq(30,198, by=24), right = FALSE)
h$counts = cumsum(h$counts)


#To plot the cumulative histogram
plot(h, main=‘Cumulative Histogram’, xlab=‘Survival Time’,
ylab=‘Cumulative Frequency’, col=‘grey’)


Below, we show the histograms obtained by using the above R code.
```



Another graph called the *ogive curve*, which represents the *cumulative frequency distribution* (c.d.f.), is obtained by joining the lower limit of the first bin to the upper limits of all the bins, including the last bin. Thus, the ogive curve for the data in Example 2.4.5 is as shown in Figure 2.4.11.

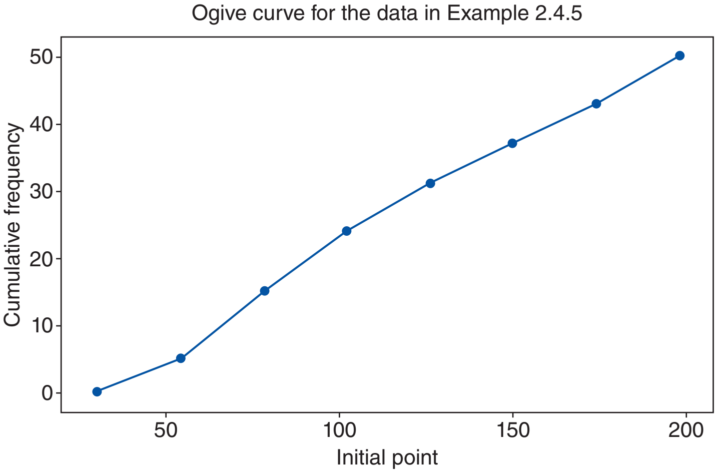


Figure 2.4.11 Ogive curve using MINITAB for the data in Example 2.4.5.

2.4.5 Line Graph

A line graph, also known as a time-series graph, is commonly used to study any trends in the variable of interest that might occur over time. In a line graph, time is marked on the horizontal axis (the x-axis) and the variable on the vertical axis (the y-axis). For illustration, we use the data of Table 2.4.4 given below in Example 2.4.6.

Example 2.4.6 (Lawn mowers) *The data in Table 2.4.4 give the number of lawn mowers sold by a garden shop over a period of 12 months of a given year. Prepare a line graph for these data.*

Table 2.4.4 Lawn mowers sold by a garden shop over a period of 12 months of a given year.

Months	January	February	March	April	May	June	July	August	September	October	November	December
LM sold	2	1	4	10	57	62	64	68	40	15	10	5

Solution: To prepare the line graph, plot the data in Table 2.4.4 using the x-axis for the months and the y-axis for the lawn mowers sold, and then join the plotted points with a freehand curve. The line graph for the data in this example is as shown in Figure 2.4.12, which was created using MINITAB (**Graph** > **Time Series Plot**).

From the line graph in Figure 2.4.12, we can see that the sale of lawn mowers is seasonal, since more mowers were sold in the summer months. Another point worth noting is that a good number of lawn mowers were sold in September when summer is winding down. This may be explained by the fact that many stores want to clear out such items as the mowing season is about to end, and many customers take advantage of clearance sales. Any mower sales during winter months may result because of a discounted price, or perhaps the store may be located where winters are very mild, and there is still a need for mowers, but at a much lower rate.

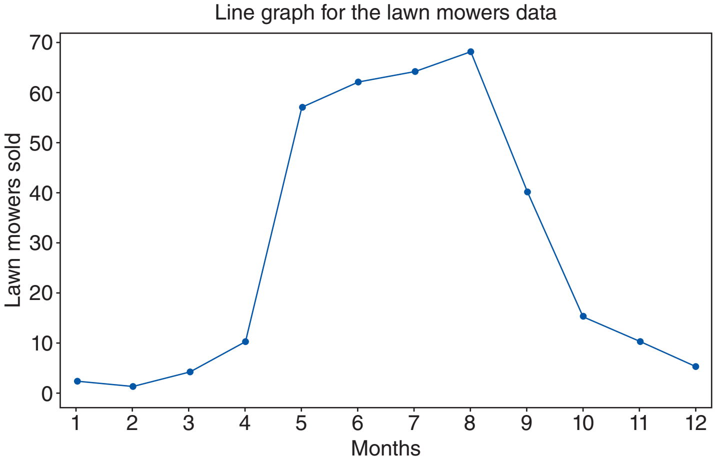


Figure 2.4.12 Line graph for the data on lawn mowers in Example 2.4.6.

2.4.6 Stem-and-Leaf Plot

Before discussing this plot, we need the concept of the *median* of a set of data. The median is the value, say M_d, that divides the data into two equal parts when the data are arranged in ascending order. A working definition is the following (a more detailed examination and discussion of the median is given in Section 2.5.1).

Definition 2.4.2 Suppose that we have a set of values, obtained by measuring a certain variable, say n times. Then, the *median* of these data, say M_d, is the value of the variable that satisfies the following two conditions:

(i) at most 50% of the values in the set are less than M_d, and
(ii) at most 50% of the values in the set are greater than M_d.

We now turn our attention to the stem-and-leaf plot invented by John Tukey. This plot is a graphical tool used to display quantitative data. Each data value is split into two parts, the part with leading digits is called the *stem*, and the rest is called the *leaf*. Thus, for example, the data value 5.15 is divided in two parts with 5 for a stem and 15 for a leaf.

A stem-and-leaf plot is a powerful tool used to summarize quantitative data. The stem-and-leaf plot has numerous advantages over both the frequency distribution table and the frequency histogram. One major advantage of the stem-and-leaf plot over the frequency distribution table is that from a frequency distribution table, we cannot retrieve the original data, whereas from a stem-and-leaf plot, we can easily retrieve the data in its original form. In other words, if we use the information from a stem-and-leaf plot, there is no loss of information, but this is not true of the frequency distribution table. We illustrate the construction of the stem-and-leaf plot with the following example.

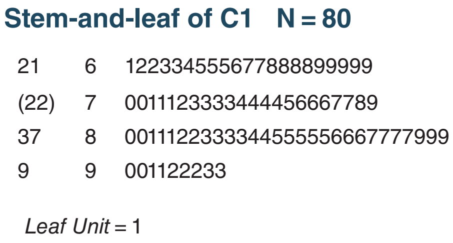
Example 2.4.7 (Spare parts supply) *A manufacturing company has been awarded a huge contract by the Defense Department to supply spare parts. In order to provide these parts on schedule, the company needs to hire a large number of new workers. To estimate how many workers to hire, representatives of the Human Resources Department decided to take a random sample of 80 workers and find the number of parts each worker produces per week. The data collected is given in Table 2.4.5. Prepare a stem-and-leaf diagram for these data.*

Table 2.4.5 Number of parts produced per week by each worker.

73	70	68	79	84	85	77	75	61	69	74	80	83	82	86	87	78	81	68	71
74	73	69	68	87	85	86	87	89	90	92	71	93	67	66	65	68	73	72	83
76	74	89	86	91	92	65	64	62	67	63	69	73	69	71	76	77	84	83	85
81	87	93	92	81	80	70	63	65	62	69	74	76	83	85	91	89	90	85	82

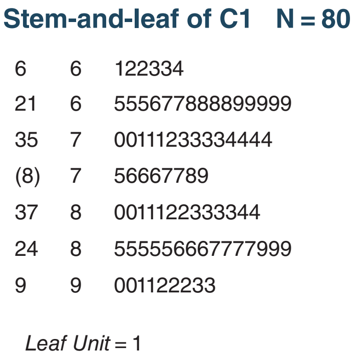
Solution: The stem-and-leaf plot for the data in Table 2.4.5 is as shown in Figure 2.4.13.

The first column in Figure 2.4.13 gives the cumulative frequency *starting from the top and from the bottom of the column* but ending at the stem that lies before the stem containing the median. The number in parentheses indicates the stem that contains the median value of the data, and the frequency of that stem.



```
Stem-and-leaf of C1   N = 80

21     6   122334555677888899999
(22)   7   0011123333444456667789
37     8   0011122333344555556667777999
9      9   001122233

Leaf Unit = 1
```

Figure 2.4.13 Stem-and-leaf plot for the data in Example 2.4.7 with increment 10.



```
Stem-and-leaf of C1   N = 80

6      6   122334
21     6   555677888899999
35     7   00111233334444
(8)    7   56667789
37     8   0011122333344
24     8   555556667777999
9      9   001122233

Leaf Unit = 1
```

Figure 2.4.14 Stem-and-leaf plot for the data in Example 2.4.7 with increment 5.

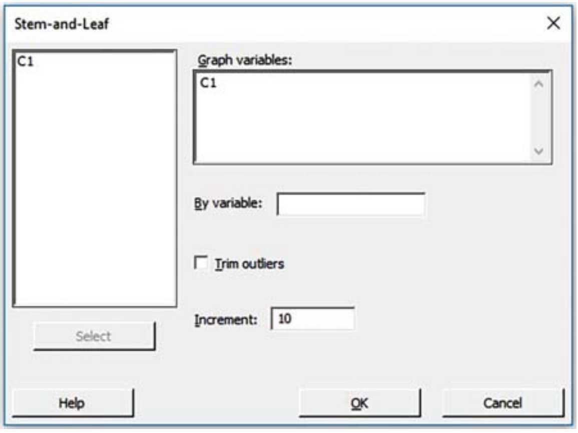
Carefully examining the stem-and-leaf plot in Figure 2.4.13, we note that the data are clustered together; each stem has many leaves. This situation is the same as when we have too few classes in a frequency distribution table. Thus having too many leaves on the stems makes the stem-and-leaf diagram less informative. This problem can be resolved by splitting each stem into two, five, or more stems depending on the size of the data. Figure 2.4.14 shows a stem-and-leaf plot when we split each stem into two stems.

The first column in the above stem-and-leaf plots counts from the top, and at the bottom is the number of workers who have produced up to and beyond certain number of parts. For example, in Figure 2.4.14, the entry in the third row from the top indicates that 35 workers produced fewer than 75 parts/wk, whereas the entry in the third row from the bottom indicates that 37 workers produced at least 80 parts/wk. The number within parentheses gives the number of observations on that stem and indicates that the *middle* value or the median of the data falls on that stem. Furthermore, the stem-and-leaf plots in Figure 2.4.14 is more informative than Figure 2.4.13. For example, the stem-and-leaf plot in Figure 2.4.14 clearly indicates that the data is bimodal, whereas Figure 2.4.13 fails to provide this information. By rotating the stem-and-leaf plot counterclockwise through

90°, we see that the plot can serve the same purpose as a histogram, with stems as classes or bins, leaves as class frequencies, and columns of leaves as rectangles or bars. Unlike the frequency distribution table and histogram, the stem-and-leaf plot can be used to answer questions such as, "What percentage of workers produced between 75 and 83 parts (inclusive)?" Using the stem-and-leaf plot, we readily see that 20 of 80, or 25% of the workers, produced between 75 and 83 parts (inclusive). However, using the frequency distribution table, this question cannot be answered, since the interval 80–85 cannot be broken down to get the number of workers producing between 80 and 83 per week. It is clear that we can easily retrieve the original data from the stem-and-leaf plot.

MINITAB

1. Enter the data in column C1.
2. From the Menu bar, select **Graph** > **Stem-and-leaf**. This prompts the following dialog box to appear on the screen. In this dialog box,



Stem-and-Leaf

C1

Graph variables: C1

By variable:

□ Trim outliers

Increment: 10

3. Enter C1 in the box under **Graph variables**.
4. Enter the desired increment in the box next to **Increment**. For example, in Figures 2.4.13 and 2.4.14, we used increments of 10 and 5, respectively.
5. Click **OK**. The stem-and-leaf plot will appear in the Session window.

USING R

We can use the built in 'stem()' function in R to generate stem-and-leaf plots. The extra argument 'scale' can be used to define the length of the stem. The task can be completed by running the following R code in R Console window.

```
SpareParts = c(73,70,68,79,84,85,77,75,61,69,74,80,83,82,86,87,78,81,
68,71,74,73,69,68,87,85,86,87,89, 90,92,71,93,67,66,65,68,73,72,83,
76,74,89,86,91,92,65,64,62,67,63,69,73,69,71,76,77,84,83,85,81,87,
93,92,81,80,70,63,65,62,69,74,76,83,85,91,89,90,85,82)
```

```
#To plot stem-and-leaf plot
stem(SpareParts, scale = 1)

#R output
                    The decimal point is 1 digit(s) to the right of the |
                      6  |    122334
                      6  |    555677888899999
                      7  |    00111233334444
                      7  |    56667789
                      8  |    0011122333344
                      8  |    555556667777999
                      9  |    001122233
```

PRACTICE PROBLEMS FOR SECTION 2.4

1. Prepare a pie chart and bar chart for the data in Problem 2 of Section 2.3.
2. Prepare a pie chart and bar chart for the data in Problem 3 of Section 2.3 and comment on the cars the senior citizens like to drive.
3. Prepare a line graph for the data in Problem 5 of Section 2.3 and state whether these data show any patterns. Read the data columnwise.
4. Use the data in Problem 6 of Section 2.3 to do the following:
   (a) Construct a frequency histogram for these data.
   (b) Construct a relative frequency histogram for these data.
   (c) Construct a frequency polygon for these data.
   (d) Construct an ogive curve for these data.
5. Construct two stem-and-leaf diagrams for the data in Problem 4 of Section 2.3, using increments of 10 and 5, and comment on which diagram is more informative.
6. Construct a stem-and-leaf diagram for the data in Problem 6 of Section 2.3. Then, reconstruct the stem-and-leaf diagram you just made by dividing each stem into two stems and comment on which diagram is more informative.
7. A manufacturing company is very training oriented. Every month the company sends some of its engineers for six-sigma training. The following data give the number of engineers who were sent for six-sigma training during the past 30 months:

18	20	16	30	14	16	22	24	16	14	16	19	18	24	23
28	18	12	18	15	17	21	22	25	27	23	19	18	20	26

   Using technology, prepare a complete frequency distribution table for these data.
8. A manufacturer of men's shirts is interested in finding the percentage of cotton in fabric used for shirts that are in greater demand. In order to achieve her goal, she took a random sample of 30 men who bought shirts from a targeted market. The

following data shows the cotton content of shirts bought by these men (some men bought more than one shirt, so that here $n = 88 > 30$):

35	25	65	35	50	35	40	50	65	55	25	55	65	35	25	35	45	55	65	55	35	45
35	45	20	35	40	45	35	65	35	50	35	30	35	65	35	25	35	20	35	65	35	30
35	65	35	30	25	35	65	35	65	35	20	35	25	35	30	35	65	35	65	35	30	35
30	65	35	30	35	20	35	65	35	55	35	30	35	65	35	65	35	30	35	65	35	35

(a) Prepare a single-valued frequency distribution table for these data.

(b) Prepare a pie chart for these data and comment on the cotton contents in these shirts.

9. The following data give the number of patients treated per day during the month of August at an outpatient clinic in a small California town:

20	30	25	35	32	46	40	38	44	41	37	35	40	41	43	38
37	35	32	40	23	26	27	29	21	23	28	33	39	20	29	

(a) Prepare a complete frequency distribution table for the data using six classes.

(b) On how many days during August were 36 or more patients treated in the clinic?

10. The following data give the number of parts that do not meet certain specifications in 50 consecutive batches manufactured in a given plant of a company:

16	19	22	25	27	18	36	30	20	24	29	40	30	31	34	36	21
25	24	28	26	30	24	16	19	21	30	24	20	22	24	32	27	18
24	20	17	33	35	29	32	36	39	28	26	17	18	25	27	29	

Construct a frequency histogram and a frequency polygon for these data.

11. A manufacturer of a part is interested in finding the life span of the part. A random sample of 30 parts gave the following life spans (in months):

23	25	30	32	36	42	28	24	21	43	46	48	39	30	34
35	24	21	16	54	25	34	37	23	24	28	26	19	27	37

Construct a relative frequency histogram and a cumulative frequency histogram for these data. Comment on the life span of the part in question.

12. The following data give the number of accidents per week in a manufacturing plant during a period of 25 weeks:

0	4	1	2	3	4	2	1	0	3	4	0	2	1	3	2	4	2	0	5	3	5	0	1	4

(a) Construct a single-valued frequency distribution table for these data.
(b) Construct a frequency histogram for these data.
(c) During how many weeks was the number of accidents less than 2?
(d) During how many weeks was the number of accidents at least 3?
(e) What is the relative frequency of 0 accidents?

13. Compressive strengths were measured on 60 samples of a new metal that a car manufacturing company is considering for use in bumpers with better shock-absorbent properties. The data are shown below:

59.7	58.3	59.0	61.5	58.7	63.8	68.2	65.6	63.5	62.4	59.4	63.2	64.5	60.0	60.5
61.5	68.5	66.6	61.3	58.5	59.2	61.3	60.4	60.6	62.1	63.5	64.4	67.3	67.9	64.2
65.4	69.3	67.3	64.5	62.3	71.7	60.7	60.2	66.7	68.5	64.2	65.1	67.0	59.5	61.7
63.1	67.5	68.5	69.2	61.5	62.3	68.4	66.5	65.7	69.3	62.5	68.0	60.5	62.3	60.5

(a) Prepare a complete frequency distribution table.
(b) Construct a frequency histogram.
(c) Construct a relative frequency histogram.
(d) Construct a frequency and relative frequency polygon.
(e) Construct a cumulative frequency histogram and then draw the ogive curve for these data.

14. Refer to the data in Problem 13 above. Construct a stem-and-leaf diagram for these data.

15. The following data give the consumption of electricity in kilowatt-hours during a given month in 30 rural households in Maine:

260	290	280	240	250	230	310	305	264	286	262	241	209	226	278
206	217	247	268	207	226	247	250	260	264	233	213	265	206	225

(a) Construct, using technology, a stem-and-leaf diagram for these data.
(b) Comment on what you learn from these data.

2.5 NUMERICAL MEASURES OF QUANTITATIVE DATA

Methods used to derive *numerical measures* for sample data as well as population data are known as *numerical methods.*

Definition 2.5.1 Numerical measures computed by using data of the *entire population* are referred to as *parameters.*

Definition 2.5.2 Numerical measures computed by using *sample data* are referred to as *statistics.*

In the field of statistics, it is standard practice to denote *parameters* by letters of the Greek alphabet and *statistics* by letters of the Roman alphabet.

We divide *numerical measures* into three categories: (i) measures of centrality, (ii) measures of dispersion, and (iii) measures of relative position. Measures of centrality give us information about the center of the data, measures of dispersion give information about the variation around the center of the data, and measures of relative position tell us what percentage of the data falls below or above a given measure.

2.5.1 Measures of Centrality

Measures of centrality are also known as measures of central tendency. Whether referring to measures of centrality or central tendency, the following measures are of primary importance:

1. Mean
2. Median
3. Mode

The mean, also sometimes referred to as the arithmetic mean, is the most useful and most commonly used measure of centrality. The median is the second most used, and the mode is the least used measure of centrality.

Mean

The mean of a sample or a population is calculated by dividing the sum of the data measurements by the number of measurements in the data. The sample mean is also known as sample average and is denoted by $\bar{X}$ (read as X bar), and the population mean is denoted by the Greek letter μ (read as meu). These terms are defined as follows:

$$\text{Population mean : } \quad \mu = (X_1 + X_2 + \cdots + X_N)/N = \sum_{i=1}^{N} X_i/N \tag{2.5.1}$$

$$\text{Sample mean : } \bar{X} = (X_1 + X_2 + \cdots + X_n)/n = \sum_{i=1}^{n} X_i/n \tag{2.5.2}$$

In (2.5.1), X_i denotes the value of the variable X possessed by the ith member of the population, $i = 1, 2, \ldots, N$. In (2.5.2), the X_i denotes the ith measurement made in a sample of size n. Here, N and n denote the population and sample size, respectively, and $n < N$. The symbol $\sum$ (read as sigma) denotes the summation over all the measurements. Note that here $\bar{X}$ is a statistic, and μ is a parameter.

Example 2.5.1 (Workers' hourly wages) *The data in this example give the hourly wages (in dollars) of randomly selected workers in a manufacturing company:*

8, 6, 9, 10, 8, 7, 11, 9, 8

Find the sample average and thereby estimate the mean hourly wage of these workers.

Solution: Since wages listed in these data are for only some of the workers in the company, the data represent a sample. Thus, we have $n = 9$, and the observed $\sum_{i=1}^{9} X_i$ is

$$\sum_{i=1}^{n} X_i = (8 + 6 + 9 + 10 + 8 + 7 + 11 + 9 + 8) = 76$$

Thus, the sample average is observed to be

$$\bar{X} = \sum_{i=1}^{n} X_i/n = 76/9 = 8.44$$

In this example, the average hourly wage of these employees is \$8.44 an hour.

Example 2.5.2 (Ages of employees) *The following data give the ages of all the employees in a city hardware store:*

22, 25, 26, 36, 26, 29, 26, 26

Find the mean age of the employees in that hardware store.

Solution: Since the data give the ages of *all* the employees of the hardware store, we are dealing with a *population.* Thus, we have

$$N = 8, \quad \sum_{i=1}^{N} X_i = (22 + 25 + 26 + 36 + 26 + 29 + 26 + 26) = 216$$

so that the population mean is

$$\mu = \sum_{i=1}^{N} X_i/N = 216/8 = 27$$

In this example, the mean age of the employees in the hardware store is 27 years.

Even though the formulas for calculating sample average and population mean are very similar, it is important to make a clear distinction between the *sample mean* or *sample average* $\bar{X}$ and the *population mean* μ for all application purposes.

Sometimes, a data set may include a few observations that are quite small or very large. For examples, the salaries of a group of engineers in a big corporation may include the salary of its CEO, who also happens to be an engineer and whose salary is much larger than that of other engineers in the group. In such cases, where there are some very small and/or very large observations, these values are referred to as *extreme values* or *outliers.* If extreme values are present in the data set, then the mean is not an appropriate measure of centrality. Note that any extreme values, large or small, adversely affect the mean value. In such cases, the median is a better measure of centrality since the median is unaffected by a few extreme values. Next, we discuss the method to calculate the median of a data set.

Median

We denote the median of a data set by M_d. To determine the median of a data set of size n, we take the following steps:

Step 1. Arrange the observations in the data set in an ascending order and rank them from 1 to n.

Step 2. Find the rank of the median that is given by

$$\boxed{\text{Rank} = \begin{cases} (n+1)/2 & \text{if } n \text{ odd} \\ n/2 \quad \text{and} \quad n/2+1 & \text{if } n \text{ even} \end{cases} \qquad (2.5.3)}$$

We can check manually that the conditions of Definition 2.4.2 are satisfied.

Step 3. Find the value of the observation corresponding to the rank of the median found in (2.5.3). If $x_{(i)}$ denotes the ith largest value in the sample, and if

(i) n odd, say $n = 2m+1$, then the median is $x_{(m+1)}$
(ii) n even, say $n = 2m$, then the median is taken as $(x_{(m)} + x_{(m+1)})/2$

Note that in the second case, we take median as the average of $x_{(m)}$ and $x_{(m+1)}$ because both satisfy the two conditions of Definition 2.4.2, resulting in their mean being adopted as a compromise between these two values for the value of M_d.

We now give examples of each case, n odd and n even.

Example 2.5.3 (Alignment pins for the case of n odd, $n = 11$) *The following data give the length (in mm) of an alignment pin for a printer shaft in a batch of production:*

30, 24, 34, 28, 32, 35, 29, 26, 36, 30, 33

Find the median alignment pin length.

Solution:

Step 1. Write the data in an ascending order and rank them from 1 to 11, since $n = 11$.

Observations in ascending order	24	26	28	29	30	30	32	33	34	35	36
Ranks	1	2	3	4	5	6	7	8	9	10	11

Step 2. Rank of the median = $(n+1)/2 = (11+1)/2 = 6$.

Step 3. Find the value corresponding to rank 6, which in this case is equal to 30. Thus, the median alignment pin length is $M_d = 30$ mm. This means that at most 50% alignment pins in the sample are of length less than or equal to 30 *and* at the most 50% are of length greater than or equal to 30 mm.

Example 2.5.4 (Sales data) *For the case of n even (i.e., $n = 16$), the following data describe the sales (in thousands of dollars) for 16 randomly selected sales personnel distributed throughout the United States:*

10 8 15 12 17 7 20 19 22 25 16 15 18 250 300 12

Find the median sale of these individuals.

Solution:

Step 1. Write the data in an ascending order and rank them from 1 to 16, since $n = 16$.

Observations in ascending order	7	8	10	12	12	15	15	16	17	18	19	20	22	25	250	300
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Step 2. Rank of the median = $(16 + 1)/2 = 8.5$.

Step 3. Following our previous discussion, the median in this case is the average of the values corresponding to their ranks of 8 and 9. Thus, the median of this data is $M_d = (16 + 17)/2 = 16.5$. In other words, the median sales of the given individuals is $16,500. We remark that eight observations fall below 16.5, and eight observations fall above 16.5.

It is important to note that the median may or may not be one of the values of the data set as in this case. Whenever the sample size is odd, the median is the center value, and whenever it is even, the median is always the average of the two middle values when the data are arranged in the ascending order.

Finally, note that the data in this example contain the two values $250,000 and $300,000. These large values seem to be the sales of top-performing sales personnel and may be considered as outliers. In this case, the mean of these data is

$$\begin{aligned}\bar{X} &= (7 + 8 + 10 + 12 + 12 + 15 + 15 + 16 + 17 + 18 + 19 + 20 + 22 + 25 + 250 + 300)/16\\ &= 47.875\end{aligned}$$

Note that the mean of 47.875 is much larger than the median of 16.5. It is obvious that the mean of these data has been adversely affected by the outliers. Hence, in this case, the mean does not adequately represent the measure of centrality of the data set, so that the median would more accurately identify the location of the center of the data.

Furthermore, if we replace the extreme values of 250 and 300, for example, by 25 and 30, respectively, then the median will not change, whereas the mean becomes 16.937, namely $16,937. Thus, the new data obtained by replacing the values 250 and 300 with 25 and 30, respectively, do not contain any outliers. The new mean value is more consistent with the true average sales.

Weighted Mean

Sometimes, we are interested in finding the sample average of a data set where each observation is given a relative importance expressed numerically by a set of values called weights. We illustrate the concept of weighted mean with the following example.

Example 2.5.5 (GPA data) *Elizabeth took five courses in a given semester with 5, 4, 3, 3, and 2 credit hours. The grade points she earned in these courses at the end of the semester were 3.7, 4.0, 3.3, 3.7, and 4.0, respectively. Find her GPA for that semester.*

Solution: Note that in this example, the data points 3.7, 4.0, 3.3, 3.7, and 4.0 have different weights attached to them; that is, the weights are the credit hours for each course. Thus, to find Elizabeth's GPA, we cannot simply find the arithmetic mean. Rather, in this case, we need to find the mean called the *weighted mean*, which is defined as

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^n w_iX_i}{\sum_{i=1}^n w_i} \tag{2.5.4}$$

where $w_1, w_2, \ldots, w_n$ are the weights attached to $X_1, X_2, \ldots, X_n$, respectively. Thus, her GPA is given by

$$\bar{X}_w = \frac{5(3.7) + 4(4.0) + 3(3.3) + 3(3.7) + 2(4.0)}{5 + 4 + 3 + 3 + 2} = 3.735$$

Mode

The mode of a data set is the value that occurs most frequently. The mode is the least used measure of centrality. When items are produced via mass production, for example, clothes of certain sizes or rods of certain lengths, the modal value is of great interest. Note that in any data set, there may be no mode, or conversely, there may be multiple modes. We denote the mode of a data set by M_0.

Example 2.5.6 (Finding a mode) *Find the mode for the following data set:*

3, 8, 5, 6, 10, 17, 19, 20, 3, 2, 11

Solution: In the data set of this example, each value occurs once except 3, which occurs twice. Thus, the mode for this set is

$$M_0 = 3$$

Example 2.5.7 (Data set with no mode) *Find the mode for the following data set:*

1, 7, 19, 23, 11, 12, 1, 12, 19, 7, 11, 23

Solution: Note that in this data set, each value occurs twice. Thus, this data set does not have any mode.

Example 2.5.8 (Tri-modal data set) *Find the mode for the following data set:*

5, 7, 12, 13, 14, 21, 7, 21, 23, 26, 5

Solution: In this data set, values 5, 7, and 21 occur twice, and the rest of the values occur only once. Thus, in this example, there are three modes, that is,

$$M_0 = 5, \quad 7, \quad \text{and} \quad 21$$

These examples show that there is no mathematical relationship among the mean, mode, and median in the sense that if we know any one or two of these measures (i.e., mean, median, or mode), then we cannot find the missing measure(s) without using the data values. However, the values of mean, mode, and median do provide important information about the type or shape of the frequency distribution of the data. Although the shape of the frequency distribution of a data set could be of any type, in practice, the most frequently encountered distributions are the three types shown in Figure 2.5.1. The location of the measures of centrality as shown in Figure 2.5.1 provides the information about the shape of the frequency distribution of a given data.

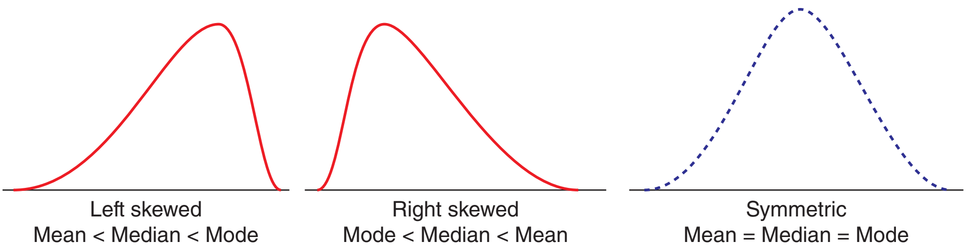


Figure 2.5.1 Frequency distributions showing the shape and location of measures of centrality.

Definition 2.5.3 A data set is *symmetric* when the values in the data set that lie equidistant from the mean, on either side, occur with equal frequency.

Definition 2.5.4 A data set is *left-skewed* when values in the data set that are greater than the median occur with relatively higher frequency than those values that are smaller than the median. The values smaller than the median are scattered to the left far from the median.

Definition 2.5.5 A data set is *right-skewed* when values in the data set that are smaller than the median occur with relatively higher frequency than those values that are greater than the median. The values greater than the median are scattered to the right far from the median.

2.5.2 Measures of Dispersion

In the previous section, we discussed measures of centrality, which provide information about the location of the center of frequency distributions of the data sets under consideration. For example, consider the frequency distribution curves shown in Figure 2.5.2. Measures of central tendency do not portray the whole picture of any data set. For example, it can be seen in Figure 2.5.2 that the two frequency distributions have the same mean, median, and mode. Interestingly, however, the two distributions are very

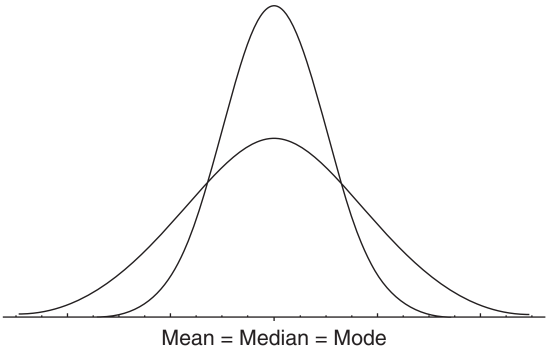


Figure 2.5.2 Two frequency distribution curves with equal mean, median, and mode values.

different. The major difference is in the variation among the values associated with each distribution. It is important, then, for us to know about the variation among the values of the data set. Information about variation is provided by measures known as *measures of dispersion*. In this section, we study three measures of dispersion: *range*, *variance*, and *standard deviation*.

Range

The range of a data set is the easiest measure of dispersion to calculate. Range is defined as

$$\text{Range} = \text{Largest value} - \text{Smallest value} \tag{2.5.5}$$

The range is not an efficient measure of dispersion because it takes into consideration only the largest and the smallest values and none of the remaining observations. For example, if a data set has 100 distinct observations, it uses only two observations and ignores the remaining 98 observations. As a rule of thumb, if the data set contains 10 or fewer observations, the range is considered a reasonably good measure of dispersion. For data sets containing more than 10 observations, the range is not considered to be an efficient measure of dispersion.

Example 2.5.9 (Tensile strength) *The following data gives the tensile strength (in psi) of a sample of certain material submitted for inspection. Find the range for this data set:*

8538.24, 8450.16, 8494.27, 8317.34, 8443.99, 8368.04, 8368.94, 8424.41, 8427.34, 8517.64

Solution: The largest and the smallest values in the data set are 8538.24 and 8317.34, respectively. Therefore, the range for this data set is

$$\text{Range} = 8538.24 - 8317.34 = 220.90$$

Variance

One of the most interesting pieces of information associated with any data is how the values in the data set vary from one another. Of course, the range can give us some idea of variability. Unfortunately, the range does not help us understand centrality. To better understand variability, we rely on more powerful indicators such as the *variance*, which is a value that focuses on how far the observations within a data set deviate from their mean.

For example, if the values in a data set are $X_1, X_2, \ldots, X_n$, and the sample average is $\bar{X}$, then $X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$ are the deviations from the sample average. It is then natural to find the sum of these deviations and to argue that if this sum is large, the values differ too much from each other, but if this sum is small, they do not differ from each other too much. Unfortunately, this argument does not hold, since, as is easily proved, the sum of these deviations is always zero, no matter how much the values in the data set differ. This is true because some of the deviations are positive and some are negative. To avoid the fact that this summation is zero, we can square these deviations and then take their sum. The variance is then the average value of the sum of the squared deviations from $\bar{X}$. If the data set represents a population, then the deviations are taken from the population mean μ. Thus, the *population variance*, denoted by σ^2 (read as sigma squared), is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{2.5.6}$$

Further the *sample variance*, denoted by S^2, is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.5.7}$$

For computational purposes, we give below the simplified forms for the population variance and the sample variances.

$$\text{Population variance} : \sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N}\right) \tag{2.5.8}$$

$$\text{Sample variance} : S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right) \tag{2.5.9}$$

Note that one difficulty in using the variance as the measure of dispersion is that the units for measuring the variance are not the same as those for data values. Rather, variance is expressed as a square of the units used for the data values. For example, if the data values are dollar amounts, then the variance will be expressed in squared dollars. Therefore, for application purposes, we define another measure of dispersion, called the *standard deviation*, that is directly related to the variance. We note that the standard deviation is measured in the same units as used for the data values (see (2.5.10) and (2.5.11) given below).

Standard Deviation

A standard deviation is obtained by taking the positive square root (with positive sign) of the variance. The population standard deviation σ and the sample standard deviation S are defined as follows:

$$\text{Population standard deviation} : \sigma = +\sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N}\right)} \qquad (2.5.10)$$

$$\text{Sample standard, deviation} : S = +\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)} \qquad (2.5.11)$$

Example 2.5.10 (Lengths of certain chips) *The following data give the length (in millimeters) of material chips removed during a machining operation:*

4, 2, 5, 1, 3, 6, 2, 4, 3, 5

Determine the variance and the standard deviation for these data.

Solution: There are three simple steps to calculate the variance of any data set.

Step 1. Calculate $\sum X_i$, the sum of all the data values, that is,

$$\sum X_i = 4 + 2 + 5 + 1 + 3 + 6 + 2 + 4 + 3 + 5 = 35$$

Step 2. Calculate $\sum X_i^2$, the sum of squares of all the observations, that is,

$$\sum X_i^2 = 4^2 + 2^2 + 5^2 + 1^2 + 3^2 + 6^2 + 2^2 + 4^2 + 3^2 + 5^2 = 145$$

Step 3. Since the sample size is $n = 10$, by inserting the values $\sum X_i$ and $\sum X_i^2$, calculated in Step 1 and Step 2 in formula (2.5.9), the sample variance is given by

$$S^2 = \frac{1}{10-1}\left(145 - \frac{(35)^2}{10}\right) = \frac{1}{9}(145 - 122.5) = 2.5$$

The standard deviation is obtained by taking the square root of the variance, that is

$$S = \sqrt{2.5} = 1.58$$

Note: It is important to remember the value of S^2, and therefore of S, is always greater than zero, except when all the data values are equal, in which case $S^2 = 0$.

Empirical Rule

We now illustrate how the standard deviation of a data set helps us measure the variability of the data. If the data have a distribution that is approximately bell-shaped, the following rule, known as the *empirical rule*, can be used to compute the percentage of data that will fall within k standard deviations from the mean $(k = 1, 2, 3)$. For the case where the data set is the set of population values, the empirical rule may be stated as follows:

1. About 68% of the data will fall within one standard deviation of the mean, that is, between $\mu - 1\sigma$ and $\mu + 1\sigma$.
2. About 95% of the data will fall within two standard deviations of the mean, that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$.
3. About 99.7% of the data will fall within three standard deviations of the mean, that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Figure 2.5.3 illustrates these features of the empirical rule.

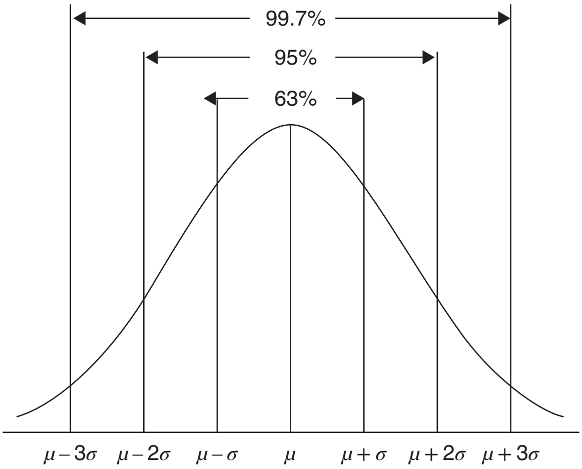


Figure 2.5.3 Application of the empirical rule.

For the case where μ and σ are unknown, the empirical rule is of the same form, but μ is replaced by $\bar{X}$ and σ replaced by S.

Example 2.5.11 (Soft drinks) *A soft-drink filling machine is used to fill 16-oz soft-drink bottles. The amount of beverage slightly varies from bottle to bottle, and it is assumed that the actual amount of beverage in the bottle forms a bell-shaped distribution with a mean 15.8 oz and standard deviation 0.15 oz. Use the empirical rule to find what percentage of bottles contain between 15.5 and 16.1 oz of beverage.*

Solution: From the information provided to us in this problem, we have $\mu = 15.8$ oz and $\sigma = 0.15$ oz. We are interested in knowing the percentage of bottles that will contain between 15.5 and 16.1 oz of beverage. We can see that $\mu \pm 2\sigma = 15.8 \pm 2(0.15) = (15.5, 16.1)$. Then comparing Figure 2.5.4 with Figure 2.5.3, it seems that approximately 95% of the bottles contain between 15.5 and 16.1 oz of the beverage, since 15.5 and 16.1 are two standard deviations away from the mean.

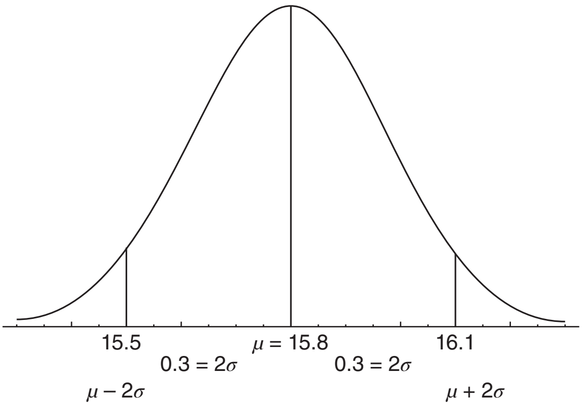


Figure 2.5.4 Distribution of amounts of soft drink contained in bottles.

Example 2.5.12 (Applying the empirical rule) *At the end of each fiscal year, a manufacturer writes off or adjusts its financial records to show the number of units of bad production occurring over all lots of production during the year. Suppose that the dollar values associated with the various units of bad production form a bell-shaped distribution with mean $\bar{X} =$ \$35,700 and standard deviation $S =$ \$2500. Find the percentage of units of bad production that has a dollar value between \$28,200 and \$43,200.*

Solution: From the information provided, we have $\bar{X} =$ \$35,700 and $S =$ \$2500. Since the limits \$28,200 and \$43,200 are three standard deviations away from the mean, applying the empirical rule shows that approximately 99.7% units of the bad production has dollar value between \$28,200 and \$43,200.

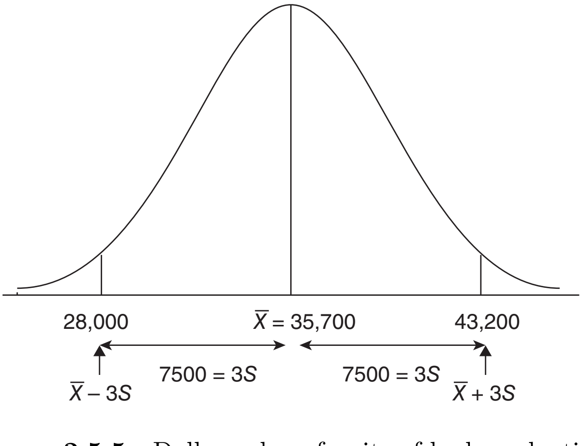


Figure 2.5.5 Dollar value of units of bad production.

If the population data have a distribution that is not bell-shaped, then we use another result, called Chebyshev's inequality, which states:

Chebyshev's inequality: For any $k > 1$, at least $(1 - 1/k^2)100\%$ of the data values fall within k standard deviations of the mean.

Figure 2.5.6a,b illustrates the basic concept of Chebyshev's inequality. Chebyshev's inequality is further discussed in Chapter 5.

The shaded area in Figure 2.5.6a contains at least $(1 - 1/k^2)100\% = (1 - 1/2^2)100\% = 75\%$ of the data values. The shaded area in Figure 2.5.6b contains at least $(1 - 1/k^2)100\% = (1 - 1/3^2)100\% \approx 89\%$ of the data values. Note that Chebyshev's inequality is also valid for sample data.

Example 2.5.13 (Using Chebyshev's inequality) *Sodium is an important component of the metabolic panel. The average sodium level for 1000 American male adults who were tested for low sodium was found to be 132 mEq/L with a standard deviation of 3 mEq/L. Using Chebyshev's inequality, determine at least how many of the adults tested have a sodium level between 124.5 and 139.5 mEq/L.*

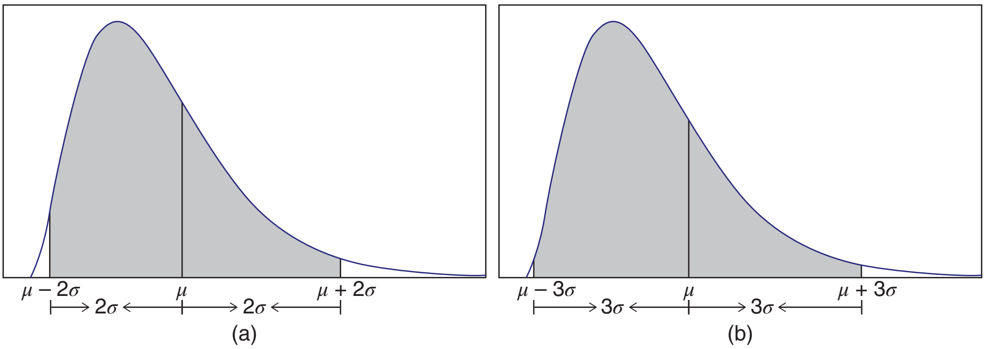


Figure 2.5.6 Shaded area lies within the intervals: (a) $[\mu - 2\sigma,\ \mu + 2\sigma]$ and (b) $[\mu - 3\sigma,\ \mu + 3\sigma]$.

Solution: From the given information, we have that the mean and the standard deviation of sodium level for these adults are

$$\bar{X} = 132 \quad \text{and} \quad S = 3$$

To find how many of 1000 adults have their sodium level between 124.5 and 139.5 mEq/L, we need to determine the value of k. Since each of these values is 7.5 points away from the mean, then using Chebyshev's inequality, the value of k is such that $kS = 7.5$, so that

$$k = 7.5/3 = 2.5.$$

Hence, the number of adults in the sample who have their sodium level between 124.5 and 139.5 mEq/L is at least

$$(1 - 1/(2.5)^2) \times 1000 = (1 - 0.16) \times 1000 = 840$$

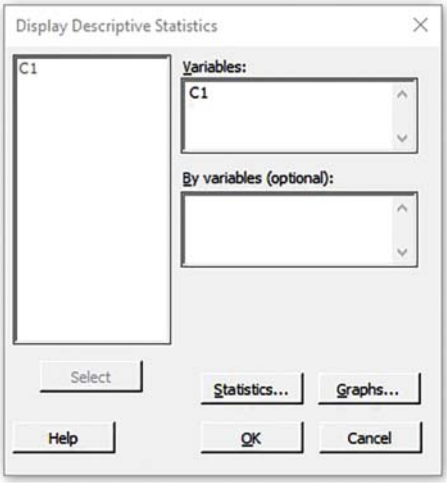
Numerical measures can easily be determined by using any one of the statistical packages discussed in this book. We illustrate the use of MINITAB and R with the following example. The use of JMP is discussed in Section 2.11, which is available on the book website: www.wiley.com/college/gupta/statistics2e.

Example 2.5.14 (Using MINITAB and R) *Calculate numerical measures for the following sample data:*

6, 8, 12, 9, 14, 18, 17, 23, 21, 23

MINITAB

1. Enter the data in column C1.
2. From the Menu bar, select **Stat** > **Basic Statistics** > **Display Descriptive Statistics**. This prompts the following dialog box to appear on the screen:



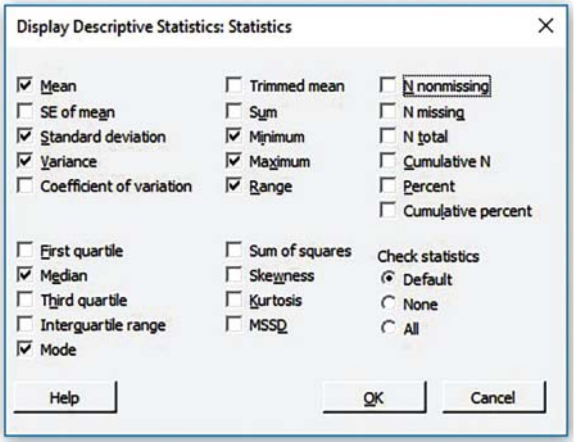
Display Descriptive Statistics

C1

Variables: C1

By variables (optional):

3. In this dialog box, enter C1 in the box under variables and click at the box **Statistics**.... Then, the following dialog box appears:

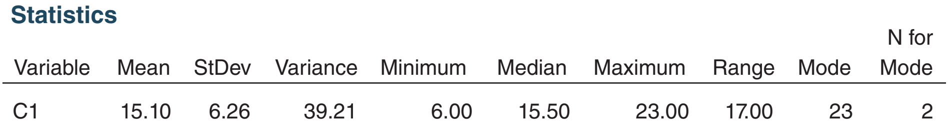


Display Descriptive Statistics: Statistics

☑ Mean
□ SE of mean
☑ Standard deviation
☑ Variance
□ Coefficient of variation
□ First quartile
☑ Median
□ Third quartile
□ Interquartile range
☑ Mode
□ Trimmed mean
□ Sum
☑ Minimum
☑ Maximum
☑ Range
□ Sum of squares
□ Skewness
□ Kurtosis
□ MSSD
□ N nonmissing
□ N missing
□ N total
□ Cumulative N
□ Percent
□ Cumulative percent

Check statistics
● Default
○ None
○ All

In this dialog box, check the statistics you would like to determine (for instance, we checked Mean, Mode, Median, Variance, Standard Deviation, Minimum, Maximum, and Range) and then click **OK**, again, click **OK**. The numerical measures shown below appear in the Session window:



Statistics

									N for
Variable	Mean	StDev	Variance	Minimum	Median	Maximum	Range	Mode	Mode
C1	15.10	6.26	39.21	6.00	15.50	23.00	17.00	23	2

USING R:

We can use the built in ‘summary()’ function in R to get basic summary statistics. However, the extra functions ‘mean()’, ‘sd()’, ‘var()’, and ‘median()’ are used to calculate the sample mean, standard deviation, variance, and median, respectively. The mode can be obtained using the manual calculation specify in the following R code. The task can be completed by running the following R code in the R Console window.

```
data = c(6, 8, 12, 9, 14, 18, 17, 23, 21, 23)

#To obtain summary statistics
summary(data)

                 Min.   1st Qu.   Median   Mean    3rd Qu.   Max.
                 6.00   9.75      15.50    15.10   20.25     23.00

#To obtain the mean, median, standard deviation, and variance
mean(data)
[1] 15.1
median(data)
[1] 15.5
sd(data)
[1] 6.261878
var(data)
[1] 39.21111

# To obtain the mode
names(table(data))[table(data) == max(table(data))]
[1] "23"
```

PRACTICE PROBLEMS FOR SECTION 2.5

1. The data given below gives the viscosity of paper pulp measured over a period of 30 days:

117.9	117.7	121.9	116.8	118.9	121.2	119.0	117.5	120.1	122.6
120.1	124.1	120.1	118.4	117.2	121.7	122.2	122.0	121.2	120.4
119.8	121.6	118.1	119.3	121.1	119.6	117.9	119.4	120.8	122.1

   (a) Determine the mean, median, and mode for these data.
   (b) Determine the standard deviation for this sample data set.
   (c) Use the results of part (a) to comment on the shape of the data.

2. Use the values of the mean ($\bar{X}$) and the standard deviation (S) found in Problem 1 to determine the number of data points that fall in the intervals $(\bar{X} - S,\ \bar{X} + S)$, $(\bar{X} - 2S,\ \bar{X} + 2S)$, and $(\bar{X} - 3S,\ \bar{X} + 3S)$. Assuming that the distribution of this data set is bell-shaped, use the empirical rule to find the number of data points that you would expect to fall in these intervals. Compare the two results and comment.

3. Reconsider the data in Problem 4 of Section 2.3, reproduced below:

25	30	26	26	25	16	21	22	21	27	15	24	19	20	24	16	30	28
24	23	15	15	21	28	18	15	21	27	26	28	17	19	24	26	27	17
27	19	22	27	16	25	16	30	18									

   (a) Determine the mean and median for these data.
   (b) Determine the standard deviation for these data.
   (c) Determine what percentage of the data fall within 2.5 standard deviations of the mean.

4. Reconsider the data in Problem 5 of Section 2.3, reproduced here:

148	167	171	177	175	165	134	177	168	142
126	166	130	122	157	138	163	129	143	145
141	162	147	141	164	137	149	146	132	157

   (a) Determine the mean and median for these data.
   (b) Determine the range, variance, and the standard deviation for these sample data.
   (c) Determine what percentage of the data fall within two standard deviations of the mean.

5. Reconsider the data in Problem 6 of Section 2.3, reproduced here:

105	101	120	116	108	112	118	119	107	100
107	120	113	113	101	102	102	100	101	100
118	106	114	100	104	101	107	113	110	100
109	108	100	104	110	113	118	100	119	120

   (a) Determine the mean, median, and mode for these data.
   (b) Determine the range, variance, and the standard deviation, for these data.

6. Use the values of the mean ($\bar{X}$) and the standard deviation (S) found in part (a) of Problem 5 above to determine the number of data points that fall in the intervals, $(\bar{X} - S, \bar{X} + S)$, $(\bar{X} - 2S, \bar{X} + 2S)$, and $(\bar{X} - 3S, \bar{X} + 3S)$. Assuming that the distribution of this data set is bell-shaped, use the empirical rule to find the number of data points that you would expect to fall in these intervals. Compare the two results and comment.

7. John is a very hard-working and an ambitious student. In a certain semester, he took in fact six courses that had 5, 4, 4, 3, 3, and 2 credit hours. The grade points he earned in these courses at the end of the semester were 3.7, 4.0, 3.3, 4.0, 3.7, and 4.0, respectively. Find his GPA for that semester.

8. The following data shows the tread depth in millimeters (mm) of 20 of tires selected randomly from a large shipment received by a dealer:

6.28	7.06	6.50	6.76	6.82	6.92	6.86	7.15	6.57	6.48
6.64	6.94	6.49	7.14	7.16	7.10	7.08	6.48	6.40	6.54

   (a) Find the mean and the median for these data.
   (b) Find the variance and standard deviation for these data.
   (c) If the desired tread depth on these tires is 7 mm, what you can say about the quality of these tires?

9. The average salary of engineers in a manufacturing company is \$55,600 with a standard deviation of \$4500. Assuming that the shape of the distribution of salaries is bell-shaped, estimate the ranges of salaries within which approximately 68% and 95% of all the engineers' salaries are expected to fall.

10. According to Chebyshev's inequality, what we can say about the lower limit of the percentage of any set of data values that must lie within k standard deviations of the mean when (a) $k = 3$, (b) $k = 3.5$, (c) $k = 4$, (d) $k = 5$?

11. Consider the following data giving the lengths (to the nearest centimeter) of a part used in the fuselage of a plane:

24	22	23	25	22	21	23	24	20	22
22	24	21	23	23	20	22	24	23	25

   (a) Determine the mean ($\bar{X}$) and the standard deviation (S) of these data.
   (b) Calculate the intervals $(\bar{X} \pm S)$, $(\bar{X} \pm 2S)$, and $(\bar{X} \pm 3S)$.
   (c) Determine the percentage of parts whose length lie within two and three standard deviations of the mean. Use these percentages to verify if the Chebyshev's inequality is valid.

2.6 NUMERICAL MEASURES OF GROUPED DATA

A set of data presented in the form of a frequency distribution table is called *grouped data*. So far, in this chapter, we have learned how to compute measures of centrality and measures of dispersion for ungrouped data. In this section, we will learn how to compute these measures for a grouped data.

Definition 2.6.1 The average of the lower and upper limits of a class (bin) is called the *class midpoint* or *class mark*.

To compute the measures of centrality and dispersion for a grouped data, each measurement in a given class is approximated by its midpoint. Thus, the measures computed from grouped data are only approximate values of measurements obtained from the original data. The actual approximation, of course, depends on the class width. Thus, in certain cases, the approximate values may be very close to the actual values, and in other cases, they may be very far apart. A word of caution: measurements obtained from grouped data should only be used when it is not possible to retrieve the original data.

2.6.1 Mean of a Grouped Data

In order to compute the average of a grouped data set, the first step is to find the midpoint (m) of each class, which is defined as

$$m = (\text{Lower limit} + \text{Upper limit})/2$$

Then, the population mean μ_G and the sample average $\bar{X}_G$ are defined as follows:

$$\mu_G = \sum f_i m_i / N \tag{2.6.1}$$

$$\bar{X}_G = \sum f_i m_i / n \tag{2.6.2}$$

Here, summation is over the number of classes involved, m_i = midpoint of the ith class, f_i = frequency of the ith class, that is, the number of values falling in the ith class, N = population size, and n = sample size.

Example 2.6.1 (Ages of basketball fans) *Find the mean of the grouped data that is the frequency distribution of a group of 40 basketball fans watching a basketball game (see Table 2.6.1).*

Solution: Using formula (2.6.2), we have

$$\bar{X}_G = \sum f_i m_i / n = 1350/40 = 33.75$$

Table 2.6.1 Age distribution of a group of 40 basketball fans watching a basketball game.

Class	Frequency (f_i)	Class midpoint (m_i)	$f_i \times m_i$
$[10, 20)$	8	(10+20)/2 = 15	120
$[20, 30)$	10	(20+30)/2 = 25	250
$[30, 40)$	6	(30+40)/2 = 35	210
$[40, 50)$	11	(40+50)/2 = 45	495
$[50, 60]$	5	(50+60)/2 = 55	275

2.6.2 Median of a Grouped Data

To compute the median M_G of a grouped data set, follow the steps given below:

Step 1. Determine the rank of the median that is given by

$$\text{Rank of } M_G = (n+1)/2$$

Step 2. Locate the class containing the median and then proceed as follows:
Add the frequencies of classes starting from class 1 and continue until the sum becomes greater than or equal to $(n+1)/2$. Then, the class containing the median is identified.

Step 3. Once we identify the class containing the rank of the median, then the median is given by

$$M_G = L + (c/f)w \tag{2.6.3}$$

where

L = lower limit of the class containing the median
$c = (n+1)/2$ – (sum of the frequencies of all classes preceding the class containing the median)
f = frequency of the class containing the median
w = class width of the class containing the median

Example 2.6.2 (Median of a grouped data) *Find the median of the grouped data in Example 2.6.1.*

Solution:

Step 1. Rank of the median $= (40+1)/2 = 20.5$.

Step 2. Add the frequencies until the sum becomes greater than or equal to 20.5, that is,

$$8 + 10 + 6 = 24 > 20.5$$

Stop at the class whose frequency is 6, so that the class containing the median is $[30, 40)$.

Step 3. Using (2.6.3), we have

$$M_G = 30 + \{[20.5 - (8+10)]/6\}10 = 30 + (2.5/6)10 = 34.17$$

2.6.3 Mode of a Grouped Data

To find the mode of a grouped data set is simple. This is because we need only to find the class with the highest frequency. The mode of the grouped data is equal to the midpoint of that class. But, if there are several classes with the same highest frequency, then there are several modes that are equal to the midpoints of such classes.

In Example 2.6.1, the mode is equal to the midpoint of the class $[40, 50)$, since it has the highest frequency, 11. Thus

$$\text{Mode} = (40 + 50)/2 = 45$$

2.6.4 Variance of a Grouped Data

The population and the sample variance of grouped data are computed by using the following formulas:

$$\text{Population variance} : \sigma_G^2 = \frac{1}{N}\left[\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{N}\right] \tag{2.6.4}$$

$$\text{Sample variance} : S_G^2 = \frac{1}{n-1}\left[\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{n}\right] \tag{2.6.5}$$

where f, m, N, and n are as defined earlier in this section.

Example 2.6.3 (Variance of a grouped data) *Determine the variance of the grouped data in Example 2.6.1.*

Solution: From the data in Table 2.6.1, we have

$$\sum f_i m_i = 8(15) + 10(25) + 6(35) + 11(45) + 5(55) = 1350$$

$$\sum f_i m_i^2 = 8(15)^2 + 10(25)^2 + 6(35)^2 + 11(45)^2 + 5(55)^2 = 52{,}800$$

Substituting these values and $n = 40$ in formula (2.6.5), we obtain

$$S_G^2 = \frac{1}{40-1}[52{,}800 - (1350)^2/40] = \frac{1}{39}[52{,}800 - 45, 562.5] = \frac{1}{39}[7237.5] = 185.577$$

The population and the sample standard deviation are found by taking the square root of the corresponding variances. For example, the standard deviation for the grouped data in Example 2.6.1 is

$$S_G = \sqrt{185.577} = 13.62$$

PRACTICE PROBLEMS FOR SECTION 2.6

1. Use the frequency distribution table you prepared in Problem 4 of Section 2.3 to do the following:
   (a) Determine the mean, median, and mode of the grouped data.
   (b) Determine the variance and the standard deviation of the grouped data.

2. Use the frequency distribution table you prepared in Problem 5 of Section 2.3, to do the following:
   (a) Determine the mean, median, and mode of the grouped data.
   (b) Determine the variance and the standard deviation of the grouped data.
3. Use the frequency distribution table you prepared in Problem 6 of Section 2.3 to do the following:
   (a) Determine the mean, median, and mode of the grouped data.
   (b) Determine the variance and the standard deviation of the grouped data.
4. The following data give the systolic blood pressures of 30 US male adults whose ages are 30–40 years old:

113	122	111	119	125	113	123	122	115	115
112	117	121	116	118	116	109	109	112	116
122	109	110	115	109	115	120	122	125	111

   (a) Determine the mean, median, and mode of these data.
   (b) Determine the variance and the standard deviation of these data.
   (c) Prepare a frequency distribution table for these data.
   (d) Use the frequency distribution table of part (c) to determine the mean, median, and mode of the grouped data. Compare your results with those in part (a) and comment.
   (e) Use the frequency distribution table of part (c) to determine the variance and the standard deviation of the grouped data. Compare your results with those in part (b) and comment.
5. The data below gives the time (in minutes) taken by 36 technicians to complete a small project:

55	58	46	58	49	46	41	60	59	41	59	43
42	40	44	42	58	46	58	58	40	51	59	49
48	46	42	43	56	48	41	54	56	57	48	43

   Construct a frequency distribution table for these data. Find the mean and the standard deviation of the grouped data, and then compare them with the actual mean and standard deviation (that is, the ungrouped $\bar{X}$ and S) of these data.

2.7 MEASURES OF RELATIVE POSITION

This section introduces measures of relative position that divide the data into percentages to help locate any data value in the whole data set. Commonly used measures of relative position are percentiles and quartiles: *percentiles* divide the data into one hundred parts, such that each part contains at the most 1% of the data, and *quartiles* divide the data into four parts, such that each part contains at the most 25% of the data. Then from quartiles, we can derive another measure, which is called the *interquartile range* (IQR), to give the range of the middle 50% of the data values. This is obtained by first organizing the data in an ascending order and then trimming 25% of the data values from the lower and the upper ends. A *quantile* is a value which divide a distribution or an ordered sample such that a specified proportion of observations fall below that value. For instance, the *percentiles* and *quartiles* are very specific quantiles.

2.7.1 Percentiles

Percentiles divide the data into one hundred equal parts; each part contains at the most 1% of the data and is numbered from 1 to 99. For example, the median of a data set is the 50th percentile, which divides the data into two equal parts so that at most 50% of the data fall below the median and at most 50% of the data fall above it. The procedure for determining the percentiles is similar to the procedure used for determining the median. We compute the percentiles as follows:

Step 1. Write the data values in an ascending order and rank them from 1 to n.
Step 2. Find the rank of the pth percentile ($p = 1, 2, \ldots, 99$), which is given by

$$\text{Rank of the } p\text{th percentile} = p \times [(n+1)/100] \qquad (2.7.1)$$

Step 3. Find the data value that corresponds to the rank of the pth percentile.

We illustrate this procedure with the following example.

Example 2.7.1 (Engineers' salaries) *The following data give the salaries (in thousands of dollars) of 15 engineers in a corporation:*

62 48 52 63 85 51 95 76 72 51 69 73 58 55 54

(a) Find the 70th percentile for these data.
(b) Find the percentile corresponding to the salary of $60,000.

Solution: (a) We proceed as follows:

Step 1. Write the data values in the ascending order and rank them from 1 to 15.

Salaries	48	51	51	52	54	55	58	62	63	69	72	73	76	85	95
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Step 2. Find the rank of the 70th percentile, which from (2.7.1) is given by

$$70 \times ((15+1)/100) = 11.2$$

Step 3. Find the data value that corresponds to the ranks 11 and 12, which in this example are 72 and 73, respectively. Then, the 70th percentile is given by

$$70\text{th percentile} = 72 + (73 - 72)(0.2) = 72.2$$

Thus, the 70th percentile of the salary data is $72,200.
That is, at most 70% of the engineers are making less than $72,200 and at most 30% of the engineers are making more than $72,200.

(b) Now we want to find the percentile p corresponding to a given value x. This can be done by using the following formula:

$$p = \frac{\text{Number of data values} \leq x}{n+1} \times 100 \qquad (2.7.2)$$

Thus, the percentile corresponding to the salary of $60,000 is

$$p = \left(\frac{7}{15+1}\right) \times 100 = 43.75 \approx 44$$

Hence, the engineer who makes a salary of $60,000 is at the 44th percentile. In other words, at most 44% of the engineers are making less than $60,000, or at most 56% are making more than $60,000.

2.7.2 Quartiles

In the previous discussion, we considered the percentiles that divide the data into 100 equal parts. Some of the percentiles have special importance, such as the 25th, 50th, and 75th percentiles. These percentiles are also known as the first, second, and third quartiles (denoted by Q_1, Q_2, and Q_3), respectively. Sometimes, they are also known as the lower, middle, and upper quartiles, respectively. The second quartile is the same as the median. Thus, to determine the values of the different quartiles, one has to simply find the 25th, 50th, and 75th percentiles (see Figure 2.7.1).

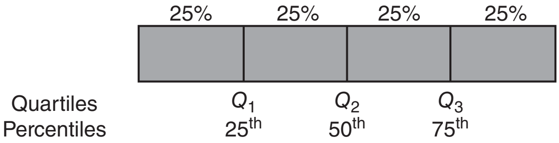


Figure 2.7.1 Quartiles and percentiles.

2.7.3 Interquartile Range (IQR)

Often we are more interested in finding information about the middle 50% of a population. A measure of dispersion relative to the middle 50% of the population or sample data is known as the *IQR*. This range is obtained by trimming 25% of the values from the bottom and 25% of the values from the top. This is equivalent to finding the spread between the first quartile and the third quartile, which is IQR and is defined as

$$\text{IQR} = Q_3 - Q_1 \tag{2.7.3}$$

Example 2.7.2 (Engineers' salaries) *Find the IQR for the salary data in Example 2.7.1:*

Salaries: 48, 51, 51, 52, 54, 55, 58, 62, 63, 69, 72, 73, 76, 85, 95

Solution: In order to find the IQR, we need to find the quartiles Q_1 and Q_3 or equivalently 25th percentile and the 75th percentile. We can easily see that the ranks of 25th and 75th percentile are given by (see (2.7.1))

$$\text{Rank of 25th percentile} = 25 \times [(15+1)/100] = 4$$

$$\text{Rank of 75th percentile} = 75 \times [(15+1)/100] = 12$$

Consulting Step 3 of Example 2.7.1, we find that $Q_1 = 52$ (52 has rank 4) and $Q_3 = 73$ (73 has rank 12). Thus, the middle 50% of the engineers earn between \$52,000 and \$73,000. The IQR in this example is given by

$$\text{IQR} = \$73{,}000 - \$52{,}000 = \$21{,}000$$

Notes:

1. The IQR gives an estimate of the range of the middle 50% of the population.
2. The IQR is potentially a more meaningful measure of dispersion than the range as it is not affected by the extreme values that may be present in the data. By trimming 25% of the data from the bottom and 25% from the top, we eliminate the extreme values that may be present in the data set. Thus, the IQR is often used as a measure of comparison between two or more data sets on similar studies.

2.7.4 Coefficient of Variation

The coefficient of variation is usually denoted by *cv* and is defined as the ratio of the standard deviation to the mean expressed as a percentage:

$$cv = \frac{\text{Standard deviation}}{|\text{Mean}|} = \frac{S}{|\bar{X}|} \times 100\% \qquad (2.7.4)$$

where $|\bar{X}|$ is the absolute value of the mean. The coefficient of variation is a relative comparison of a standard deviation to its mean and is unitless. The *cv* is commonly used to compare the variability in two populations. For example, we might want to compare the disparity of earnings for technicians who have the same employer but work in two different countries. In this case, we would compare the coefficient of variation of the two populations rather than compare the variances, which would be an invalid comparison. The population with a greater coefficient of variation, generally speaking, has more variability than the other. As an illustration, we consider the following example.

Example 2.7.3 *A company uses two measuring instruments, one to measure the diameters of ball bearings and the other to measure the length of rods it manufactures. The quality control department of the company wants to find which instrument measures with more precision. To achieve this goal, a quality control engineer takes several measurements of a ball bearing by using one instrument and finds the sample average* $\bar{X}$ *and the standard deviation* S *to be 3.84 and 0.02 mm, respectively. Then, he/she takes several measurements of a rod by using the other instrument and finds the sample average* $\bar{X}$ *and the standard deviation* S *to be 29.5 and 0.035 cm, respectively. Estimate the coefficient of variation from the two sets of measurements.*

Solution: Using formula (2.7.4), we have

$$cv_1 = (0.02/3.84)100\% = 0.52\%$$

$$cv_2 = (0.035/29.5)100\% = 0.119\%$$

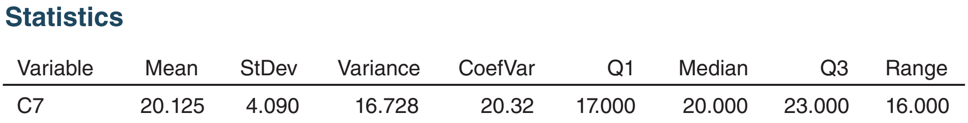
The measurements of the lengths of rod are relatively less variable than of the diameters of the ball bearings. Therefore, we can say the data show that instrument 2 is more precise than instrument 1.

Example 2.7.4 (Bus riders) *The following data gives the number of persons who take a bus during the off-peak time schedule (3–4 pm.) from Grand Central to Lower Manhattan in New York City. Using technology, find the numerical measures for these data:*

17	12	12	14	15	16	16	16	16	17	17	18	18	18	19	19	20	20	20	20
20	20	20	20	21	21	21	22	22	23	23	23	24	24	25	26	26	28	28	28

MINITAB

1. Enter the data in column C1.
2. From the Menu bar, select **Stat** > **Basic Statistics** > **Display Descriptive Statistics**:
3. In the dialog box that appears, enter C1 under **variables** and select the option **Statistics** . . . .
4. A new dialog box **Descriptive Statistics: Statistics** appears. In this dialog box, select the desired statistics and click **OK** in the two dialog boxes. The values of all the desired statistics as shown below will appear in the Session window.



Statistics

Variable	Mean	StDev	Variance	CoefVar	Q1	Median	Q3	Range
C7	20.125	4.090	16.728	20.32	17.000	20.000	23.000	16.000

USING R

We can use a few built in functions in R to get basic summary statistics. Functions ‘mean()’, ‘sd()’, and ‘var()’ are used to calculate the sample mean, standard deviation, and variance, respectively. The coefficient of variation can be calculated manually using the mean and variance results. The ‘quantile()’ function is used to obtain three quantiles and the minimum and maximum. The function ‘range()’ as shown below can be used to calculate the range of the data. The task can be completed by running the following R code in the R Console window.

```
x = c(17,12,12,14,15,16,16,16,16,17,17,18,18,18,19,19,20,20,20,20,20,
20,20,20,21,21,21,22,22,23,23,23, 24,24,25,26,26,28,28,28)

#To concatenate resulting mean, standard deviation, variance, and coefficient of variation
c(mean(x), sd(x), var(x), 100*sd(x)/mean(x))
[1] 20.125000 4.089934 16.727564 20.322656
```

```
#To obtain quartiles including min and max
quantile(x)
                          0%   25%   50%   75%   100%
                          ---------------------------
                          12   17    20    23    28

#To obtain the range we find Max-Min
range(x)[2]-range(x)[1]
[1] 16
```

2.8 BOX-WHISKER PLOT

In the above discussion, several times we made a mention of extreme values. At some point we may wish to know what values in a data set are extreme values, also known as *outliers*. An important tool called the *box-whisker plot* or simply *box plot*, and invented by J. Tukey, helps us answer this question. Figure 2.8.1 illustrates the construction of a box plot for any data set.

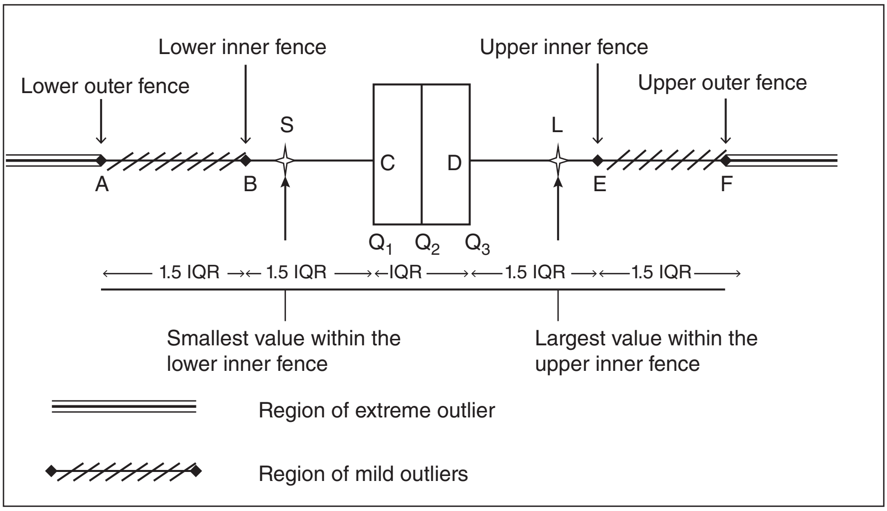


Figure 2.8.1 Box-whisker plot.

2.8.1 Construction of a Box Plot

Step 1. Find the quartiles Q_1, Q_2, and Q_3 for the given data set.

Step 2. Draw a box with its outer lines of the box standing at the first quartile (Q_1) and the third quartile (Q_3), and then draw a line at the second quartile (Q_2). The line at Q_2 divides the box into two boxes, which may or may not be of equal size.

Step 3. From the outer lines, draw straight lines extending outwardly up to three times the IQR and mark them as shown in Figure 2.8.1. Note that each distance between the points A and B, B and C, D and E, and E and F is equal to one and a one-half times distance between the points C and D, or one and one-half times IQR. The points S and L are, respectively, the smallest and largest data points that fall within the inner fences. The lines from S to C and D to L are called the whiskers.

2.8.2 How to Use the Box Plot

About the Outliers

1. Any data points that fall beyond the lower and upper outer fences are the extreme outliers. These points are usually excluded from the analysis.
2. Any data points between the inner and outer fences are the mild outliers. These points are excluded from the analysis only if we were convinced that these points are somehow recorded or measured in error.

About the Shape of the Distribution

1. If the second quartile (median) is close to the center of the box and each of the whiskers is approximately of equal length, then the distribution is *symmetric.*
2. If the right box is substantially larger than the left box and/or the right whisker is much longer than the left whisker, then the distribution is *right-skewed.*
3. If the left box is substantially larger than the right box and/or the left whisker is much longer than the right whisker, then the distribution is *left-skewed.*

Example 2.8.1 (Noise levels) *The following data gives the noise level measured in decibels (a usual conversation by humans produces a noise level of about 75 dB) produced by 15 different machines in a very large manufacturing plant:*

85 80 88 95 115 110 105 104 89 87 96 140 75 79 99

Construct a box plot and examine whether the data set contains any outliers.

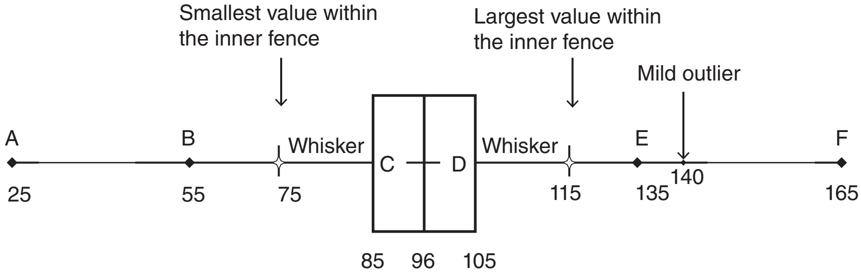


Figure 2.8.2 Box plot for the data in Example 2.8.1.

Solution: First we arrange the data in the ascending order and rank them.

Data values	75	79	80	85	88	89	95	96	97	99	104	105	110	115	140
Ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

We then find the ranks of the quartiles Q_1, Q_2, and Q_3. Thus, we have ($n = 15$)

Rank of Q_1 = (25/100)(15+1) = 4
Rank of Q_2 = (50/100)(15+1) = 8
Rank of Q_3 = (75/100)(15+1) = 12

Therefore, the values of Q_1, Q_2, and Q_3 are 85, 96, and 105, respectively, and the interquartile range is

$$\text{IQR} = Q_3 - Q_1 = 105 - 85 = 20$$

$$(1.5) \times \text{IQR} = (1.5) \times 20 = 30$$

Figure 2.8.2 shows the box plot for these data. Figure 2.8.2 shows that the data include one outlier. In this case, action should be taken to modify the machine that produced a noise level of 140 dB. Use MINITAB to create a box plot:

MINITAB

1. Enter the data in column C1.
2. From the Menu bar, select **Graph** > **Boxplot**. This prompts a dialog box to appear. In this dialog box, select **simple** and click **OK**. This prompts another dialog box to appear.
3. In this dialog box, enter C1 in the box under the **Graph variables** and click **OK**. Then, the box plot shown in Figure 2.8.3 will appear in the Session window.

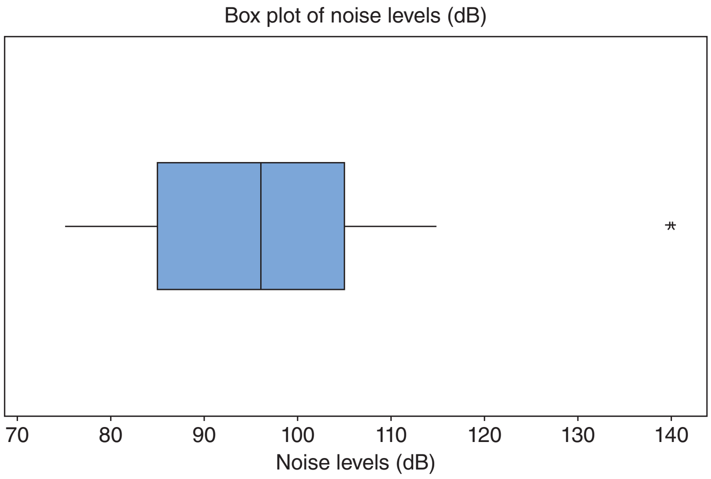


Figure 2.8.3 MNITAB printout of box plot for the data in Example 2.8.1.

USING R

We can use a built in ‘boxplot()’ function in R to generate box plots. Extra arguments such as inserting a heading, labeling y-axis, and coloring can be done as shown in the following R code.

```
NoiseLevels = c(75,79,80,85,88,89,95,96,97,99,104,105,110,115,140)

#To plot boxplot
boxplot(NoiseLevels, main = 'Box plot of Noise Levels (dB)',
ylab = 'Noise Levels (dB)', col = 'grey')
```

Example 2.8.2 (Bus riders') *From the bus riders' data in Example 2.7.4, we have*

12	12	14	15	16	16	16	16	17	17	17	18	18	18	19	19	20	20	20	20
20	20	20	20	21	21	21	22	22	23	23	23	24	24	25	26	26	28	28	28

(a) Find the mean, mode, and median for these data.
(b) Prepare the box plot for the data.
(c) Using the results of parts (a) and (b), verify if the data are symmetric or skewed. Examine whether the conclusions made using the two methods, the results of part (a) and (b) about the shapes of the distribution, are the same or not.
(d) Using the box plot, check if the data contain any outliers.
(e) If in part (c) the conclusion is that the data are symmetric, then find the standard deviation and verify if the empirical rule holds or not.

Solution: The sample size in this problem is $n = 40$. Thus, we have

(a) Mean $\bar{X} = \sum X_i/n = 800/40 = 20$, mode $= 20$, and median $= 20$
(b) To prepare the box plot, we first find the quartiles Q_1, Q_2, and Q_3.
Rank of $Q_1 = (25/100)(40 + 1) = 10.25$ Rank of $Q_2 = (50/100)(40 + 1) = 20.5$ Rank of $Q_3 = (75/100)(40 + 1) = 30.75$.

Since the data presented in this problem are already in the ascending order, we can easily see that the quartiles Q_1, Q_2, and Q_3 are

$$Q_1 = 17, Q_2 = 20, \textit{ and } Q_3 = 23$$

The interquartile range is IQR $= Q_3 - Q_1 = 23 - 17 = 6$. Thus, 1.5(IQR) $= 1.5(6) = 9$

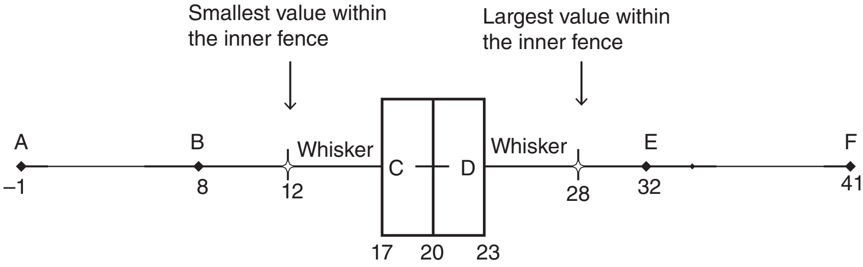


Figure 2.8.4 Box plot for the data in Example 2.8.2.

The box plot for the data is as shown in Figure 2.8.4.

(c) Both parts (a) and (b) lead to the same conclusion; that is, the data are symmetric.
(d) From the box plot in Figure 2.8.4, it is clear that the data do not contain any outliers.
(e) In part (c), we concluded that the data are symmetric, so we can proceed to calculate the standard deviation and then determine whether or not the empirical rule holds.

$$S^2 = \frac{1}{40-1}\left([12^2 + \cdots + 28^2] - \frac{(12 + \cdots + 28)^2}{40}\right) = 18.1538$$

Thus, the standard deviation is $S = 4.26$. Now it can be seen that the interval

$$(\bar{X} - S,\ \bar{X} + S) = (15.74, 24.26)$$

contains 72.5% of the data and $(\bar{X} - 2S,\ \bar{X} + 2S) = (11.48,\ 28.52)$ contains 100% of the data.

The data are slightly more clustered around the mean. But for all practical purposes, we can say that the empirical rule holds.

PRACTICE PROBLEMS FOR SECTIONS 2.7 AND 2.8

1. The following data give the amount of a beverage in 12 oz cans:

11.38	11.03	11.87	11.98	12.36	11.80	12.32	12.06	11.38	11.07
12.12	12.11	12.24	12.37	11.75	12.25	13.60	11.93	13.11	11.76
12.34	12.08	11.85	11.37	12.32	11.74	12.75	12.76	12.16	11.72
10.97	12.09	12.53	11.88	12.11	11.28	12.01	11.80	12.47	12.32

(a) Find the mean, variance, and standard deviation of these data.
(b) Find the three quartiles and the IQR for these data.
(c) Prepare a box plot for these data and determine if there are any outliers present in these data.

2. The following data gives the reaction time (in minutes) of a chemical experiment conducted by 36 chemistry majors:

55	58	46	58	49	46	41	60	59	41	59	42
40	44	42	58	46	58	58	40	51	59	48	46
42	43	56	48	41	54	56	57	48	43	49	43

(a) Find the mean, mode, and median for these data.
(b) Prepare a box plot for these data and check whether this data set contains any outliers.

3. The following data give the physics lab scores of 24 randomly selected of physics majors:

21	18	21	18	20	18	18	59	19	20	20	20
19	18	21	58	19	22	19	18	22	18	22	56

Construct a box plot for these data and examine whether this data set contains any outliers.

4. The following data provide the number of six sigma black belt Engineers in 36 randomly selected manufacturing companies in the United States:

73	64	80	67	73	78	66	78	59	79	74	75	73	66	63	62	61	58
65	76	60	79	62	63	71	75	56	78	73	75	63	66	71	74	64	43

   (a) Find the 60th percentile of these data.
   (b) Find the 75th percentile of the data.
   (c) Determine the number of data points that fall between the 60th and 75th percentiles you found in parts (a) and (b).
   (d) Prepare the box plot for these data and comment on the shape of the data:

5. Consider the following two sets of data:

Set I																	
29	24	25	26	23	24	29	29	24	28	23	27	26	21	20			
25	24	30	28	28	29	28	22	26	30	21	26	27	25	23			
Set II																	
46	48	60	43	57	47	42	57	58	59	52	53	41	58	43	50	49	56
57	54	51	46	60	44	55	43	60	50	51	54	50	43	44	53	51	58

   (a) Find the mean and standard deviation for the two data sets.
   (b) Find the coefficient of variation for these data sets.
   (c) Determine whether one of these data sets has higher variation than the other.

6. Reconsider the data in Problem 4 of Section 2.6, and do the following:
   (a) Find the mean, variance, and standard deviation of these data.
   (b) Find the three quartiles and the IQR for these data.
   (c) Prepare a box plot for these data and determine if there are any outliers present in these data.

7. Reconsider the data in Problem 5 of Section 2.6 and do the following:
   (a) Find the mean, variance, and standard deviation of these data.
   (b) Find the three quartiles and the IQR for these data.
   (c) Prepare a box plot for these data and determine if there are any outliers present in these data.

2.9 MEASURES OF ASSOCIATION

So far in this chapter, the discussion was focused on only univariate statistics because we were interested in studying a single characteristic of a subject. In all the examples we considered, the variable of interest was either qualitative or quantitative. We now study cases involving two variables; this means examining two characteristics of a subject. The two variables of interest could be either qualitative or quantitative, but here we will consider only variables that are quantitative.

For the consideration of two variables simultaneously, the data obtained are known as *bivariate* data. In the examination of *bivariate* data, the first question is whether there is any association of interest between the two variables. One effective way to determine whether there is such an association is to prepare a graph by plotting one variable along the horizontal scale (x-axis) and the second variable along the vertical scale (y-axis). Each pair of observations (x, y) is then plotted as a point in the xy-plane. The resulting graph is called a *scatter plot.* A *scatter plot* is a very useful graphical tool because it reveals the nature and strength of associations between two variables. The following example makes the concept of association clear.

Example 2.9.1 (Cholesterol level and systolic blood pressure) *The cholesterol level and the systolic blood pressure of 10 randomly selected US males in the age group 40–50 years are given in Table 2.9.1. Construct a scatter plot of this data and determine if there is any association between the cholesterol levels and systolic blood pressures.*

Solution: Figure 2.9.1 shows the scatter plot of the data in Table 2.9.1. This scatter plot clearly indicates that there is a fairly strong upward linear trend. Also, if a straight line is drawn through the data points, then it can be seen that the data points are concentrated around the straight line within a narrow band. The upward trend indicates a positive association between the two variables, while the width of the band indicates the strength of the association, which in this case is quite strong. As the association between the two variables gets stronger and stronger, the band enclosing the plotted points becomes narrower and narrower. A downward trend indicates a negative association between the two variables.

A numerical measure of association between two numerical variables is called the *Pearson correlation coefficient*, named after the English statistician Karl Pearson (1857–1936). Note that a correlation coefficient does not measure causation. In other words, correlation and causation are different concepts. Causation causes correlation, but not necessarily the converse. The correlation coefficient between two numerical variables in a set of sample data is usually denoted by r, and the correlation coefficient for population data is denoted by the Greek letter ρ (rho). The correlation coefficient r based on n pairs of (X, Y), say $(x_i, y_i),\ i = 1, 2, \ldots, n$ is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.9.1}$$

or

$$r = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sqrt{[\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n]\,[\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n]}} \tag{2.9.2}$$

The correlation coefficient is a dimensionless measure that can attain any value in the interval $[-1, +1]$. As the strength of the association between the two variables grows, the absolute value of r approaches 1. Thus, when there is a perfect association between the two variables, $r = 1$ or -1, depending on whether the association is positive or negative. In other words, $r = 1$, if the two variables are moving in the same direction, and $r = -1$, if the two variables are moving in the opposite direction.

Table 2.9.1 Cholesterol levels and systolic BP of 10 randomly selected US males.

Subject	1	2	3	4	5	6	7	8	9	10
Cholesterol (x)	195	180	220	160	200	220	200	183	139	155
Systolic BP (y)	130	128	138	122	140	148	142	127	116	123

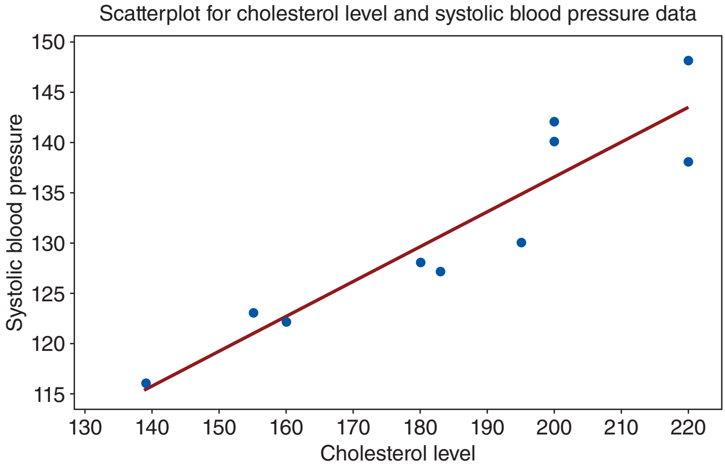
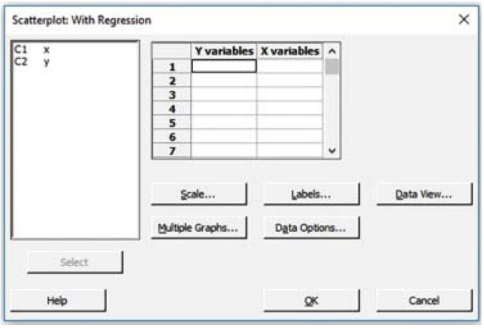


Figure 2.9.1 MINITAB printout of scatter plot for the data in Table 2.9.1.

Perfect association means that if we know the value of one variable, then the value of the other variable can be determined without any error. The other special case is when $r = 0$, which does not mean that there is no association between the two variables, but rather that there is no linear association between the two variables. As a general rule, the linear association is weak, moderate, or strong when the absolute value of r is less than 0.3, between 0.3 and 0.7, or greater than 0.7, respectively. For instance, if (2.9.1) is computed for the data in Table 2.9.1, then $r = 0.924$. Hence, we can conclude that the association between the two variables X and Y is strong.

MINITAB:

1. Enter the pairs of data in columns C1 and C2. Label the columns X and Y.
2. From the Menu bar select **Graph** > **Scatterplot** . . . . This prompts a dialog box to appear on the screen. In this dialog box, select scatterplot **With Regression** and click **OK**. This prompts the following dialog box to appear:



In this dialog box, under the X and Y variables, enter the columns in which you have placed the data. Use the desired options and click **OK**. The Scatter plot shown in Figure 2.9.1 appears in the Session window.

3. For calculating the correlation coefficient, select from the Menu bar **Stat** > **Basic Statistics** > **Correlation**. Then, enter the variables C1 and C2 in the dialog box.

USING R

We can use a built in 'plot()' function in R to generate scatter plots. Extra arguments such as 'pch' and 'cex' can be used to specify the plotting symbol and size of the symbol, respectively. Finally, the function 'abline()' can be used to embed the trend line to the scatter plot as follows. The function 'cor()' can be used to calculate the Pearson correlation coefficient. The whole task can be completed by running the following R code in the R Console window.

```
x = c(195,180,220,160,200,220,200,183,139,155)
y = c(130,128,138,122,140,148,142,127,116,123)

#To plot the data in a scatter plot
plot(x, y, pch = 20, cex = 2, main = 'Scatterplot for Cholesterol Level and Sys-
tolic Blood Pressure Data', xlab = 'Cholesterol Level', ylab = 'Systolic Blood Pres-
sure')

#To add a trend line
abline(lm(y ~ x), col = 'red')

#To calculate the Pearson correlation coefficient
cor(x, y)
[1] 0.9242063
```

The resulting R scatter plot for the data in Table 2.9.1 looks exactly the same as in the MINTAB printout in Figure 2.9.1.

PRACTICE PROBLEMS FOR SECTION 2.9

1. The following data give the heights (cm) and weights (lb) of 10 male undergraduate students:

Heights	170	167	172	171	165	170	168	172	175	172
Weights	182	172	179	172	174	179	188	168	185	169

(a) Draw a scatter plot for these data. By observing this scatter plot, do you expect the correlation between heights and weights to be positive or negative?

(b) Determine the correlation coefficient between the heights and weights.

2. The following data give the final exam scores in biology and chemistry of eight science majors:

Biology scores	85	88	78	92	89	83	79	95
Chemistry scores	90	84	86	95	94	89	84	87

(a) Draw a scatter plot for these data. By observing this scatter plot, do you expect the correlation between biology and chemistry scores to be approximately 1, −1, or 0?

(b) Determine the correlation coefficient between the biology and chemistry scores.

3. The following data show the experience (in years) and yearly salaries (in thousands of dollars) of 10 engineers:

Experience	10	12	8	15	6	11	14	16	15	12
Salaries	98	95	97	110	88	102	120	128	105	104

(a) Construct a scatter plot for these data. By observing this scatter plot, do you expect the correlation between experience and salaries to be positive or negative?

(b) Determine the correlation coefficient between the experience and salaries. Is the value of correlation coefficient consistent with what you concluded in part (a)?

4. The following scores give two managers' assessments of ten applicants for the position of a senior engineer:

Manager 1	7	8	9	7	9	8	9	7	9	6
Manager 2	8	6	9	9	8	7	9	8	7	8

(a) Construct a scatter plot for these data. By observing this scatter plot, do you expect the correlation between assessments of managers 1 and 2 to be positive or negative?

(b) Determine the correlation coefficient between the assessment managers 1 and 2. Is the value of correlation coefficient consistent with what you concluded in part (a)?

2.10 CASE STUDIES

Case Study 1 (*St. Luke's Hospital data*)[1] In the fast-paced world of health care, small fluctuations in the quarterly patient (mean) satisfaction scores may not raise any red flags. But that quickly changed for St. Luke's Hospital in Cedar Rapids, Iowa, during a leadership retreat in spring 2004. Here, managers received a shocking surprise, the effects of which are still driving improvement efforts today. The hospital's inpatient satisfaction measures, which appeared flat, had actually dipped to the 64th percentile when compared to other hospitals in the country.

People were shocked because they thought St. Luke's should be in the 80th or 90th percentiles, explains Kent Jackson, director of children's specialty services, and the leader of the hospital's patient and family experience team. "What became more significant was that in the second quarter of 2004, the hospital dropped to the 49th percentile [for inpatient satisfaction]. So, about the time that people were shocked, it was about to get worse," Jackson recalls.

[1] Source: Reproduced with permission of ASQ, Jacobson (1998).

2.10.1 About St. Luke's Hospital

St. Luke's Hospital is a fully accredited 560-bed hospital with more than 2600 associates and 381 physicians who provide 24-hour coverage in east central Iowa. Founded in 1884 as the community's first hospital, today St. Luke's is a member of the Iowa Health System. The hospital offers a comprehensive array of inpatient services including medical/surgical, cardiology, oncology, neurology, inpatient and outpatient behavioral health, and neonatal intensive care; it delivers a broad range of diagnostic and outpatient services, and provides a 24-hour trauma center and an air ambulance service. St. Luke's five-point strategy to gain patient satisfaction focused on:

- Demonstrating better quality
- Becoming the workshop of choice for physicians
- Partnering with associates
- Strengthening the core (making sure the hospital is financially sound)
- Establishing the hospital as a regional workshop of choice to better serve organizations and promote health care in the region

Tables 2.10.1–2.10.3 document the improvement at St. Luke's.

Table 2.10.1 Turnaround time for inpatient lab testing once the swarm method was implemented.

Month/Year	Inpatient tests reported within 23 minutes of receipt (%)
June 2007	63
July 2007	84
August 2007	71
September 2007	77
October 2007	84
November 2007	93
December 2007	94

Table 2.10.2 Physician satisfaction scores at St. Luke's Hospital before and after the hospital's "breaking out of the pack" strategy.

Year	Physician satisfaction score (percentile score)
2004	81
2006	95

Table 2.10.3 Inpatient satisfaction scores after St. Luke's implemented the standards of excellence of the four tools centered on the patient and family experience.

Quarter-Year	Patient satisfaction score (Percentile rank among 400–599 peer)
2nd-2004	49
3rd-2004	65
4th-2004	93
1st-2005	88
2nd-2005	93
3rd-2005	90
4th-2004	93
1st-2006	93
2nd-2006	91
3rd-2006	95
4th-2006	91
1st-2007	90

Construct the relevant graphs summarizing the data in this case study and interpret them. Then, prepare the progress report that you would like to present to the Board of Directors of the hospital.

Case Study 2 (*Part measurements using a micrometer*)[2] The goal of this study is to develop a process with reduced variation among machines producing some parts. During this study, diameters of a part are measured using a standard micrometer with readings recorded to 0.0001 of an inch. The data for this case study is available on the book website: www.wiley.com/college/gupta/statistics2e.
Do the followings:

1. Construct box plots for the three sets of data, that is, from machine-1, machine-2, and machine-3, and compare average values of the diameter of the parts produced by the three machines.
2. Determine the variances for the three sets of data, compare them, and write your conclusions.
3. Determine the coefficient of variation for the three sets of data, compare them, and write your conclusions.
4. Compare your conclusions in Parts 2 and 3 above, and comment.

2.11 USING JMP

This section is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

[2] Source: Based on data from The Engineering Statistics Handbook, National Institute of Standards and Technology (NIST).

Review Practice Problems

1. During a flu season, it is common that many workers cannot come to work because either they themselves are sick or they have to take care of their sick children. The following data give the number of employees of a company who did not come to work on 18 days during a flu season:

   7, 5, 10, 12, 6, 7, 8 10, 3, 16, 10, 9, 8, 10, 9, 8, 7, 6

   Construct a dot plot for these data. Comment on what you learn from the dot plot.

2. A saving and loan institution wants to find how many of their customers default their loan payments. The following data give the number of customers who did not make their payment on time at least once over the past 12 months:

   15, 20, 18, 16, 3, 19, 14, 17, 17, 16, 30, 15

   Construct a dot plot for these data. Comment on any patterns you observe in these data.

3. The following data give the number of machines in a shoe factory that had breakdowns during the past 21 shifts:

   3, 2, 1, 0, 2, 1, 4, 2, 0, 1, 2, 3, 1, 0, 4, 2, 1, 10, 2, 1, 2

   Construct a dot plot for these data. If you were the maintenance engineer, what would you learn from these data?

4. The following data classify a group of students who are attending a seminar on environmental issues by their class standing:

Class standing	Frequency
Freshmen	16
Sophomore	18
Junior	20
Senior	15
Graduate	30

   (a) Construct a bar chart for these data.
   (b) Construct a pie chart for these data.

5. Suppose there are two fund-raising concerts at a university. The following data give the number of students by their class standing who attended one or the other of the concerts:

Class standing	Frequency-1	Frequency-1
Freshmen	16	40
Sophomore	18	30
Junior	20	21
Senior	15	20
Graduate	30	15

(a) Construct a side-by-side bar chart for each of the concert and compare the two sets of data.
(b) Construct pie charts for each of the concerts. Do you think you can get the same information by using the two pie charts, as by using the side-by-side bar charts?

6. Refer to the data in Problem 15 of Section 2.4.
(a) Construct a frequency histogram for these data.
(b) Construct a relative-frequency histogram for these data.

7. Suppose that in a midwestern state, a legislator running for governor proposes the following state budget (in millions of dollars) for the following year:

Education	900
Medicaid	400
Other social programs	500
Road and bridges	350
Agriculture	400
Others	250

Use JMP, MINITAB, or R to do the following:
(a) Construct a bar chart for these data.
(b) Construct a pie chart for these data.
(c) Determine what percentage of the budget is used for all social programs.

8. The following data give the number of defective parts produced in 21 consecutive shifts of 1 wk

15	14	18	16	17	13	27	14	15	10	30
14	8	14	15	17	15	13	14	16	20	

(a) Prepare a line graph of these data.
(b) Check if any peaks or dips appear in the line graph.
(c) As a quality manager of the company, what would you conclude from this line graph, and what will be your line of action to reduce the number of defective parts produced?

9. Consider the following stem-and-leaf diagram:

Stem	Leaf
3	2 5 7
4	0 3 6 8 9
5	1 2 2 7 8
6	3 5 6 6 9 9
7	1 5 5 7 8

Reproduce the data set represented by the diagram.

10. The following data give the number of employees from 19 different sectors of a large company who were absent for at least two days from a certain training program:

7	5	10	12	6	7	8	10	3	16
10	9	8	10	7	6	9	11	2	

Construct a dot plot for these data and comment on what you observe in these data.

11. To improve the quality of a crucial part used in fighter jets, a quality control engineer is interested in finding the type of defects usually found in that part. He labels these defects as A, B, C, D, and E based on severity of the defect. The following data show the type of defects found in the defective parts:

B	D	A	B	C	D	B	E	B	E	D	B	C	B	E	C	D	B
E	D	B	C	B	D	B	C	D	B	A	B	C	B	D	E	B	E
B	E	C	B	D	E	B	C	E	B	E	B	C	B	D	B		

Prepare a bar chart for these data, and comment on the types of defects encountered in the parts under study.

12. The following data give the salaries (in thousands of dollars) of 62 randomly selected engineers from different manufacturing companies located in different regions of the United States:

65	45	85	68	98	95	58	62	64	54	57	58	85	120	45	56
150	140	123	65	55	66	76	88	45	50	60	66	55	46	48	98
56	66	185	56	55	77	59	67	145	166	67	58	68	69	87	89
92	85	88	77	69	76	86	81	54	145	154	190	205	85		

(a) Prepare a box whisker plot for these data.
(b) Do these data contain any mild or extreme outliers?

13. The following data give the number of cars owned by 50 randomly selected families in a metropolitan area:

3	5	2	1	2	4	3	1	2	3	4	2	3	2	5	3	1	2	4	3	2	1	2	1	4
5	1	2	3	2	3	4	2	3	1	2	3	2	4	2	3	2	1	3	1	2	4	2	3	2

(a) Construct a single-valued frequency distribution table for these data.
(b) Compute the columns of relative frequencies and percentages.
(c) Construct a bar chart for these data.
(d) What percentage of the families own at least 3 cars?
(e) What percentage of the families own at most 2 cars?

14. The following data give the total cholesterol levels (mg/100 mL) of 100 US males between 35 to 65 years of age:

177	196	150	167	175	162	195	200	167	170	179	172	176	179	177	153	177
189	185	167	151	177	191	177	175	151	173	199	167	197	188	163	174	151
183	174	177	200	182	195	160	151	177	154	150	180	170	172	153	152	194
197	192	155	174	159	193	182	175	169	180	200	194	182	188	152	196	198
171	176	200	180	161	182	188	168	165	168	160	175	193	159	183	166	198
184	172	180	195	199	156	158	152	174	151	173	166	183	194	156		

(a) Construct a frequency distribution table with classes [150, 160), [160, 170), ...
(b) What percentage of US males between 35 to 65 years of age do you estimate have cholesterol levels higher than 200 mg/100 mL?
(c) What percentage of US males between 35 to 65 years of age do you estimate have cholesterol levels less than 180 mg/100 mL?

15. We know that from a grouped data set we cannot retrieve the original data. Generate a new (hypothetical) data set from the frequency distribution table that you prepared in Problem 14. Reconstruct a frequency distribution table for the new set and comment on whether the two frequency tables should be different or not.

16. A group of dental professionals collected some data on dental health and concluded that 10% of the Americans have zero or one cavity, 50% have two or three cavities, 30% have four cavities, and rest of the 10% have five or more cavities. Construct a pie chart that describes the dental health of the American population.

17. Find the mean, median, and mode for the following sample data on credit hours for which students are registered in a given semester:

7	11	8	12	7	6	14	17	15	13

18. The following data give hourly wages of 20 workers randomly selected from a chipmaker company:

16	12	18	15	23	29	21	20	21	25
18	27	21	25	22	16	24	26	21	26

Determine the mean, median, and mode for these data. Comment on whether these data are symmetric or skewed.

19. The following data give daily sales (in gallons) of gasoline at a gas station during April:

414	450	380	360	470	400	411	465	390	384
398	412	416	454	459	395	430	439	449	453
464	450	380	398	410	399	416	426	430	425

(a) Find the mean, median, and mode for these data. Comment on whether these data are symmetric, left skewed, or right skewed.
(b) Find the range, variance, standard deviation, and the coefficient of variation for these data.

20. The owner of the gas station of Problem 19 also owns another gas station. He decided to collect similar data for the second gas station during the same period. These data are given below.

570	590	600	585	567	570	575	580	577	583
589	585	595	570	574	576	581	583	595	591
585	583	580	597	599	600	577	573	574	579

(a) Find the range, variance, standard deviation, and coefficient of variation for these data.
(b) Compare the standard deviations for the two data sets.
(c) Do you think it will be more prudent to compare the coefficients of variation rather than the two standard deviations? Why or why not?
(d) Sometimes the observations in a given data set are too large numerically to compute the standard deviation easily. However, if these observations are small, particularly when we are using paper, pen, and a small calculator, then there is little problem in computing the standard deviation. If observations are large, all one has to do is to subtract a constant from each of the data points and then find the standard deviation for the new data. The standard deviation of the new data, usually called the coded data, is exactly the same as that of the original data. Thus, for example, in Problem 20, one can subtract 567 (the smallest data point) from each data point and then find the standard deviation of the set of the coded data. Try it.

21. Collect the closing price of two stocks over a period of 10 sessions. Calculate the coefficients of variation for the two data sets and then check which stock is more risky.

22. The following data give the number of physicians who work in a hospital and are classified according to their age:

Age	[35–40)	[40–45)	[45–50)	[50–55)	[55–60)	[60–65]
Frequency	60	75	68	72	90	55

Find the mean and the standard deviation for this set of grouped data.

23. Prepare a frequency table for the data in Problem 9 of Section 2.4. Find the mean and the variance for the grouped and the ungrouped data. Then compare the values of the mean and variance of the grouped and the ungrouped data.

24. The following data give lengths (in mm) of a type of rods used in car engines.

128	118	120	124	135	130	128	116	122	120	118	125	127	123	126
124	120	132	131	119	117	124	129	131	133	115	121	122	127	127
134	128	132	135	125	120	121	126	124	123					

(a) Determine the quartiles (Q_1, Q_2, Q_3) for this data.
(b) Find the IQR for these data.
(c) Determine the value of the 70th percentile for these data.
(d) What percentage of the data falls between Q_1 and Q_3?

25. Compute $\bar{X}$, S^2, and S for the data in Problem 24. Then,
(a) Find the number of data points that fall in the intervals $\bar{X} \pm S$, $\bar{X} \pm 2S$, and $\bar{X} \pm 3S$
(b) Verify whether the empirical rule holds for these data.

26. A car manufacturer wants to achieve 35 miles/gal on a particular model. The following data give the gas mileage (rounded to the nearest mile) on 40 randomly selected brand-new cars of that model. Each car uses regular unleaded gasoline:

34	33	36	32	33	34	35	37	32	33	32	31	34	37	32	33	33	36	34	31
35	36	35	33	32	32	34	35	34	30	34	37	35	32	31	34	32	33	32	33

(a) Find the mean and the standard deviation for these data.
(b) Check whether the empirical rule holds for these data.

27. Refer to the data in Problem 26. Determine the following:
(a) The values of the three quartiles Q_1, Q_2, and Q_3.
(b) The IQR for these data.
(c) Construct a box-plot for these data and verify if the data contains any outliers.

28. The following data give the test scores of 57 students in an introductory statistics class:

68	78	92	80	87	79	74	85	86	88	91	97	71	72	81	86	60	40	76
77	20	99	80	79	89	87	87	80	83	95	92	98	87	86	95	96	75	76
79	80	85	81	77	76	84	82	83	56	68	69	91	88	69	75	74	59	61

(a) Find the values of three quartiles Q_1, Q_2, and Q_3.
(b) Find the IQR for these data.
(c) Construct the box plot for these data and check whether the data is skewed.
(d) Do these data contain any outliers?

29. The following data give the overtime wages (in dollars) earned on a particular day by a group of 40 randomly selected employees of a large manufacturing company:

30	35	45	50	25	30	36	38	42	40	46	36	30	35
24	46	42	50	40	40	35	34	34	30	28	32	30	26
28	36	40	42	40	38	38	36	45	40	36	42		

(a) Find the IQR for these data.
(b) Count what percentage of the data falling between the first and the third quartiles.
(c) Do you think the result in part (b) agrees with your expectations?

30. The following data give the time (in minutes) taken by 20 students to complete a class test:

55	63	70	58	62	71	50	70	60	65
59	62	66	71	58	70	75	70	65	68

(a) Find the mean, median, and mode for these data.
(b) Use values of the mean, median, and mode to comment on the shape of the frequency distribution of these data.

31. The following data give the yearly suggested budget (in dollars) for undergraduate books by 20 randomly selected schools from the whole United States:

690	650	800	750	675	725	700	690	650	900
850	825	910	780	860	780	850	870	750	875

(a) Find the mean and the standard deviation for these data.
(b) What percentage of schools has their budget between $\bar{X} - S$ and $\bar{X} + S$?

32. A data set has a mean of 120 and a standard deviation of 10. Using the empirical rule, find what percentage of data values fall:
(a) Between 110 and 130.
(b) Between 100 and 140.
(c) Between 90 and 150.

33. Suppose that the manager of a pulp and paper company is interested in investigating how many trees are cut daily by one of its contractors. After some investigation, the manager finds that the number of trees cut daily by that contractor forms a bell shaped distribution with mean 90 and standard deviation 8. Using the empirical rule, determine the percentage of the days he cuts
(a) Between 82 and 98 trees.
(b) Between 66 and 114 trees.
(c) More than 106 trees.
(d) Less than 74 trees.

34. The following sample data give the number of pizzas sold by a Pizza Hut over a period of 15 days:

75	45	80	90	85	90	92	86	95	95	90	86	94	99	78

(a) Prepare a box plot for these data and comment on the shape of this data set.
(b) Find the mean, median, and standard deviation of these data.

35. The following sample data give the GRE scores (actual score—2000) of 20 students who have recently applied for admission to the graduate program in an engineering school of a top-rated US university:

268	320	290	310	300	270	250	268	330	290
240	269	295	325	316	320	299	286	269	250

(a) Find the sample mean $\bar{X}$ and the sample standard deviation S.
(b) Determine the percentage of the data that falls in the interval $\bar{X} \pm 2S$.
(c) Determine the range of the middle 50% of the observations.

36. Assume that the data in Problem 35 come from a population having a bell-shaped probability distribution. Then, using the empirical rule, determine how many data values one would expect to fall within the intervals $\bar{X} \pm 2S$ and $\bar{X} \pm 3S$. Compare your results with the actual number of data values that fall in these intervals. Also, using technology, verify the assumption that the observations come from a population having a bell shaped probability distribution.

37. The following data give the number of defective parts received in the last 15 shipments at a manufacturing plant:

8	10	12	11	13	9	15	14	10	16	18	12	14	16	13

(a) Find the mean of these data.
(b) Find the standard deviation of these data.
(c) Find the coefficient of variation for these data.

38. The owner of the facility in Problem 37 has another plant where the shipments received are much larger than at the first plant. The quality engineer at this facility also decides to collect the data on defectives received in each shipment. The last 15 shipments provided the following data:

21	30	38	47	58	39	35	15	59	60	43	47	39	30	41

(a) Find the mean and the standard deviation of these data.
(b) Find the coefficient of variation of these data, compare it with the one obtained in Problem 37, and comment on which facility receives more stable shipments.

39. Prepare box plots for the data in Problems 37 and 38. Comment on the shape of the distribution of these two data sets.

40. The following data give the test scores of 40 students in a statistics class:

68	78	92	80	87	79	74	85	86	88	91	97	71	72	81	86	60	40	76	77
20	99	80	79	89	87	87	80	83	95	92	98	87	86	95	96	76	75	79	80

(a) Find the sample mean $\bar{X}$ and the sample standard deviation S for these data.
(b) Prepare a frequency distribution table for these data.
(c) Use the grouped data in part (b) to determine the grouped mean $\bar{X}_G$ and the grouped standard deviation S_G.
(d) Compare the values of $\bar{X}_G$ and S_G with the values of $\bar{X}$ and the standard deviation S. Notice that the grouped mean and grouped standard deviations are only the approximate values of the actual mean and standard deviations of the original (i.e., ungrouped) sample.

41. The following two data sets give the number of defective ball bearings found in 20 boxes randomly selected from two shipments:

Shipment I	60	65	79	71	67	68	73	56	59	63
	66	59	72	77	79	69	71	70	60	55
Shipment II	45	55	56	50	59	60	48	38	42	41
	37	57	55	49	43	39	45	51	53	55

(a) Find the quartiles for each of these two sets.
(b) Prepare the box plots for each of the two data sets and display them side by side on one sheet of graph paper.
(c) Use part (b) to compare the two shipments. Which shipment in your opinion is of better quality?

42. The following data give the number of flights that left late at a large airport over the past 30 days:

50	59	63	30	120	55	49	47	43	51	47	51	57	62	58
50	39	53	50	45	43	46	52	59	48	36	51	33	42	32

(a) Prepare a complete frequency distribution table for these data.
(b) Prepare a box plot for these data to comment on the shape of the distribution of these data. Does the set contain any outliers?
(c) Find the mean and the standard deviation for these data.

43. The following data gives the inflation rate and interest rates in the United States over 10 consecutive periods. Determine the correlation coefficient between the inflation rate and the interest rates in the United States. Interpret the value of the correlation coefficient you determined.

Period	1	2	3	4	5	6	7	8	9	10
Inflation rate	2.34	2.54	2.22	2.67	1.98	3.22	2.51	2.57	2.75	2.67
Interest rate	4.55	4.65	4.75	4.82	4.46	4.85	4.35	4.25	4.55	4.35

44. The following data gives the heights (in.) and weights (lb) of eight individuals. Determine the correlation coefficient between the heights and weights. Interpret the value of the correlation coefficient you have determined.

Individuals	1	2	3	4	5	6	7	8
Heights	77	72	73	76	72	73	77	72
Weights	156	172	195	181	158	164	164	191

45. It is generally believed that students' performance on a test is related to number of hours of sleep they have the night before the test. To verify this belief, 12 students were asked how many hours they slept on the night before the test. The following data shows the number of hours of sleep on the night before the test and the test scores of each of the 12 students. Determine the correlation coefficient between the hours of sleep and test scores. Interpret the value of the correlation coefficient you have determined.

Student	1	2	3	4	5	6	7	8	9	10	11	12
Hours of sleep	8	8	6	5	8	8	7	6	7	5	4	6
Test scores	89	84	88	85	87	97	93	90	87	90	86	72

Chapter 3

ELEMENTS OF PROBABILITY

The focus of this chapter is the study of basic concepts of probability.

Topics Covered

- Random experiments and sample spaces
- Representations of sample spaces and events using Venn diagrams
- Basic concepts of probability
- Additive and multiplicative rules of probability
- Techniques of counting sample points: permutations, combinations, and tree diagrams
- Conditional probability and Bayes's theorem
- Introducing random variables

Learning Outcomes

After studying this chapter, the reader will be able to

- Handle basic questions about probability using the definitions and appropriate counting techniques.
- Understand various characteristics and rules of probability.
- Determine probability of events and identify them as independent or dependent.
- Calculate conditional probabilities and apply Bayes's theorem for appropriate experiments.
- Understand the concept of random variable defined over a sample space.

3.1 INTRODUCTION

In day-to-day activities and decisions, we often confront two scenarios: one where we are certain about the outcome of our action and the other where we are uncertain or at a loss.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

For example, in making a decision about outcomes, an engineer knows that a computer motherboard requires four RAM chips and plans to manufacture 100 motherboards. On the one hand, the engineer is certain that he will need 400 RAM chips. On the other hand, the manufacturing process of the RAM chips produces both nondefective and defective chips. Thus, the engineer has to focus on how many defective chips could be produced at the end of a given shift and so she is dealing with uncertainty.

Probability is a measure of chance. Chance, in this context, means there is a possibility that some sort of event will occur or will not occur. For example, the manager needs to determine the probability that the manufacturing process of RAM chips will produce 10 defective chips in a given shift. In other words, one would like to measure the chance that in reality, the manufacturing process of RAM chips does produce 10 defective chips in a given shift. This small example shows that the theory of probability plays a fundamental role in dealing with problems where there is any kind of uncertainty.

3.2 RANDOM EXPERIMENTS, SAMPLE SPACES, AND EVENTS

3.2.1 Random Experiments and Sample Spaces

Inherent in any situation where the theory of probability is applicable is the notion of *performing a repetitive operation*, that is, performing a trial or experiment that is capable of being repeated over and over "under essentially the same conditions." A few examples of quite familiar repetitive operations are rolling a die, tossing two coins, drawing five screws "at random" from a box of 100 screws, dealing 13 cards from a thoroughly shuffled deck of playing cards, filling a 12-oz can with beer by an automatic filling machine, drawing a piece of steel rod, and testing it on a machine until it breaks, firing a rifle at a target 100 yards away, and burning ten 60-W bulbs with filament of type Z continuously until they all "expire."

An important feature of a repetitive operation is illustrated by the repetitive operation of firing a rifle at a 100-yard target. The shooter either hits the target or misses the target. The possible outcomes "hit" or "miss" are referred to as outcomes of the experiment "firing at a target 100 yards away." This experiment is sometimes called a *random experiment*. We will have more discussion of this at a later point. This important feature needs formalizing with the following definition.

Definition 3.2.1 In probability theory, performing a repetitive operation that results in one of the *possible* outcomes is said to be performing a *random experiment*.

One of the basic features of repetitive operations or random experiments under specified conditions is that an *outcome* may vary from trial to trial. This variation leads to the analysis of the possible outcomes that would arise if a trial were performed only once. The set of all possible outcomes under specific conditions if an experiment was performed once is called the *sample space* of the experiment and is denoted by S. It is convenient to label an outcome in a sample space S by the letter e, and call e a *sample space element* or simply an *element* or *sample point* of the sample space S. The sample space S of such elements or points is *generated* by the operations or trials of a random experiment. Consider the following examples of elements or sample points that constitute a sample space.

Example 3.2.1 (Rolling a die) *If a die is rolled once, the sample space S thus generated consists of six possible outcomes; that is, the die can turn up faces numbered 1, 2, 3, 4, 5, or 6. Thus, in this case,*

$$S = \{1, 2, 3, 4, 5, 6\}$$

Example 3.2.2 (Tossing two coins) *If two coins are tossed, say a nickel and a dime, and if we designate head and tail on a nickel by* H *and* T, *respectively, and head and tail on a dime by* h *and* t, *respectively, the sample space* S *generated by tossing the two coins consists of four possible outcomes. We then have that*

$$S = \{Hh, Ht, Th, Tt\}$$

As an example, Ht *denotes the outcome that the nickel, when tossed ended up showing head, while the dime, when tossed, showed tail.*

Example 3.2.3 (Sample space for item drawn using random sampling scheme) *The sample space S for drawing five screws "at random" from a box of 100 consists of all possible sets of five screws that could be drawn from 100; S contains 75,287,520 elements.*

Example 3.2.4 (Sample space for playing cards) *In dealing 13 cards from a thoroughly shuffled deck of ordinary playing cards, the sample space S consists of the 635,013,559,600 possible hands of 13 cards that could be dealt from the 52 cards of an ordinary deck.*

The sample spaces for these preceding examples are all *finite sample spaces*: they contain only a finite number of sample points. A sample space is finite as long as it contains a countable number of elements, no matter how large that number may be. For instance, in Example 3.2.4, the number of elements is very large but countable. Many problems in probability involve infinite sample spaces, that is, sample spaces containing an infinitely large number of elements that are not countable.

Example 3.2.5 (Sample space for reaction times) *A chemist studies the reaction time when a catalyst is added to a chemical at a certain temperature. In this experiment, the sample space S contains an indefinitely large number of elements when observing the reaction time.*

Example 3.2.6 (Sample space for beer refills) *The sample space S when filling a "12-oz" can with beer with an automatic filling machine under factory conditions, would contain an indefinitely large number of elements when measuring the fluid content of the filled can.*

3.2.2 Events

Suppose that S is the sample space of a random experiment that contains a finite number of outcomes or elements $e_1, e_2, \ldots, e_m$. In most probability problems, we are more interested in whether or not an outcome belongs to some set E of outcomes rather than in an individual outcome. For instance, if playing the game of craps, one is usually more interested in the total number of dots that appear when the two dice are thrown than in any particular pair of dots obtained from throwing a pair of dice. The inspector who examines five screws taken "at random" from a box of 100 is not really interested in any particular one of the 75,287,250 different sets of five screws he could have drawn; he is in fact looking for the number of defective screws he gets in the five screws he draws. In other words, he is interested in whether this outcome belongs to the set of outcomes with 0 defectives, or the set with one defective, or the set with two defectives, and so on.

Any set of outcomes in which there might be some particular interest is called an *event*. The following two examples describe two events.

Example 3.2.7 (Sample space generated by tossing two coins) *The event E of getting exactly one head in throwing the two coins of Example 3.2.2 consists of the set of two elements $\{Ht, Th\}$ from the sample space $S = \{Hh, Ht, Th, Tt\}$.*

Example 3.2.8 (Sample space for playing cards) *Suppose that 13 cards are dealt from a deck of ordinary playing cards. Such a deck has 13 cards of each of four suits, which are spades, clubs, hearts, and diamonds. As mentioned in Example 3.2.4, there are 635,013,559,600 possible hands making up the sample space for this experiment (repetitive operation). Now suppose that we are interested in the number of possible hands (elements in S) that contains exactly 12 spades. It turns out that this event (set) contains 507 elements, or put another way, there are 507 hands of 13 cards that contain exactly 12 spades out of the possible 635,013,559,600 hands when dealing 13 cards from a deck of 52 playing cards.*

Schematically, if the set of points inside the rectangle in Figure 3.2.1 represent a sample space S, we may represent an event E by the set of points inside a circle and $\bar{E}$ by the region outside the circle. Such a representation is called a *Venn diagram.*

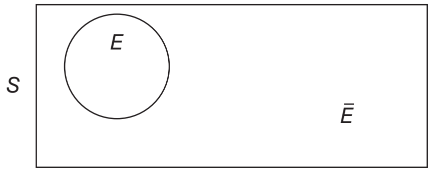


Figure 3.2.1 Venn diagram representing events E and $\bar{E}$.

Events can be described in the language of sets, and the words *set* and *event* can be used interchangeably. If E contains no elements, it is called the *empty*, *impossible*, or *null* event and is denoted by $\emptyset$. The *complement* $\bar{E}$ of an event E is the event that consists of all elements in S that are not in E. Note, again, that $\bar{E}$ is an event and that $\bar{S} = \emptyset$.

Now suppose that there are two events E and F in a sample space S. The event consisting of all elements contained in E or F, or both, is called the *union* of E and F; it is written as

$$\boxed{E \cup F \qquad (3.2.1)}$$

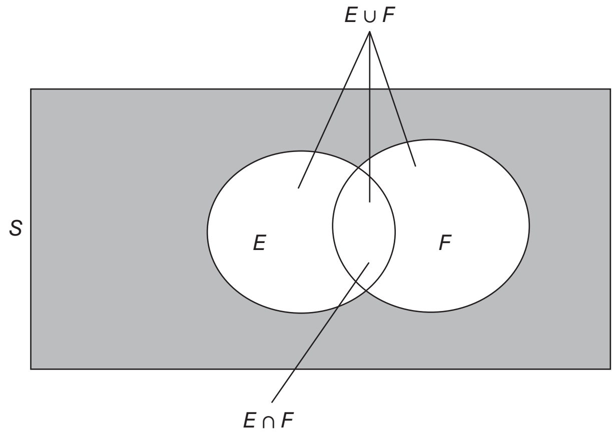


Figure 3.2.2 Venn diagram representing events $E, F, E \cup F$, and $E \cap F$.

The event consisting of all elements in a sample space S contained in both E and F is called the *intersection* of E and F; it is written as

$$E \cap F \tag{3.2.2}$$

Referring to the Venn diagram in Figure 3.2.2, note that if S is represented by the points inside the rectangle, E by the points inside the left-hand circle, and F by the points inside the right-hand circle, then $E \cup F$ is represented by the points in the region not shaded in the rectangle and $E \cap F$ is represented by the points in the region in which the two circles overlap. Also note (see Figure 3.2.1) that $E \cup \bar{E} = S$, and $E \cap \bar{E} = \emptyset$.

Example 3.2.9 (Union and intersection) *Suppose that* S *is the set of all possible hands of 13 cards,* E *is the set of all hands containing five spades, and* F *is the set of all hands containing six honor cards. An honor card is one of either a ten, Jack, Queen, King, or Ace of any suit. Then,* $E \cup F$ *is the set of all hands containing five spades* or *six honor cards,* or *both.* $E \cap F$ *is the set of all hands containing five spades* and *six honor cards.*

If there are no elements that belong to both E and F, then

$$E \cap F = \emptyset, \tag{3.2.3}$$

and the sets E and F are said to be *disjoint*, or *mutually exclusive*.

If all elements in E are also contained in F, then we say that E is a *subevent* of F, and we write

$$E \subset F \quad \text{or} \quad F \supset E \tag{3.2.4}$$

This means that if E occurs, then F necessarily occurs. We sometimes say that E is contained in F, or that F contains E, if (3.2.4) occurs.

Example 3.2.10 (Sub events) *Let* S *be the sample space obtained when five screws are drawn from a box of 100 screws of which 10 are defective. If* E *is the event consisting of all possible sets of five screws containing one defective screw and* F *is the event consisting of all possible sets of the five screws containing at least one defective, then* $E \subset F$.

If $E \subset F$ and $F \subset E$, then every element of E is an element of F, and vice versa. In this case, we say that E and F are *equal* or *equivalent* events; this is written as

$$E = F \tag{3.2.5}$$

The set of elements in E that are not contained in F is called the *difference* between E and F; this is written as

$$E - F \tag{3.2.6}$$

If F is contained in E, then $E - F$ is the *proper difference* between E and F. In this case, we have

$$\boxed{E - F = E \cap \bar{F} \qquad (3.2.7)}$$

Example 3.2.11 (Difference of two events) *If* E *is the set of all possible bridge hands with exactly five spades and if* F *is the set of all possible hands with exactly six honor cards (10, J, Q, K, A), then* $E - F$ *is the set of all hands with exactly five spades but not containing exactly six honor cards (e.g. see Figure 3.2.3).*

If $E_1, E_2, \ldots, E_k$ are several events in a sample space S, the event consisting of all elements contained in one or more of the E_i is the union of $E_1, E_2, \ldots, E_k$ written as

$$E_1 \cup E_2 \cup \cdots \cup E_k, \quad \text{or} \quad \cup_{i=1}^{k} E_i \tag{3.2.8}$$

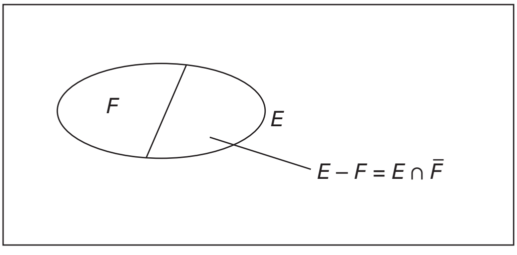


Figure 3.2.3 Venn diagram representing events $E, F, E - F$, and $E \cap \bar{F}$.

Similarly the event consisting of all elements contained in all E_i is the intersection of $E_1, E_2, \ldots, E_k$ written as

$$E_1 \cap E_2 \cap \cdots \cap E_k, \quad \text{or} \quad \cap_{i=1}^{k} E_i \tag{3.2.9}$$

If for every pair of events (E_i, E_j), $i \neq j$, from $E_1, E_2, \ldots, E_k$ we have that $E_i \cap E_j = \emptyset$, then $E_1, E_2, \ldots, E_k$ are *disjoint* and *mutually* exclusive events.

An important result concerning several events is the following theorem.

Theorem 3.2.1 *If $E_1, E_2, \ldots, E_k$ are events in a sample space* S*, then $\cup_{i=1}^{k} E_i$ and $\cap_{i=1}^{k} \bar{E}_i$ are disjoint events whose union is* S.

This result follows by noting that the events $\cup_{i=1}^{k} E_i$ and $\cap_{i=1}^{k} \bar{E}_i$ are complement of each other.

3.3 CONCEPTS OF PROBABILITY

Suppose that a sample space S, consists of a finite number, say m, of elements $e_1, e_2, \ldots, e_m$, so that the elements $e_1, e_2, \ldots, e_m$ are such that $e_i \cap e_j = \emptyset$ for all $i \neq j$ and also represent an exhaustive list of outcomes in S, so that $\cup_{i=1}^{m} e_i = S$. If the operation whose sample space is S is repeated a large number of times, some of these repetitions will result in e_1, some in e_2, and so on. (The separate repetitions are often called *trials.*) Let $f_1, f_2, \ldots, f_m$ be the fractions of the total number of trials resulting in $e_1, e_2, \ldots, e_m$, respectively. Then, $f_1, f_2, \ldots, f_m$ are all nonnegative, and their sum is 1. We may think of $f_1, f_2, \ldots, f_m$ as observed weights or measures of occurrence of $e_1, e_2, \ldots, e_m$ obtained on the basis of an experiment consisting of a large number of repeated trials. If the entire experiment is repeated, another set of f's would occur with slightly different values, and so on for further repetitions. If we think of indefinitely many repetitions, we can conceive of idealized values being obtained for the f's. It is impossible, of course, to show that in a physical experiment, the f's converge to limiting values, in a strict mathematical sense, as the number of trials increases indefinitely. So we postulate values $p(e_1), p(e_2), \ldots, p(e_m)$ corresponding to the idealized values of $f_1, f_2, \ldots, f_m$, respectively, for an indefinitely large number of trials. It is assumed that $p(e_1), p(e_2), \ldots, p(e_m)$ are all positive numbers and that

$$p(e_1) + \cdots + p(e_m) = 1 \tag{3.3.1}$$

The quantities $p(e_1), p(e_2), \ldots, p(e_m)$ are called *probabilities* of occurrence of $e_1, e_2, \ldots, e_m$, respectively.

Now suppose that E is any event in S that consists of a set of one or more e's, say $e_{i_1}, \ldots, e_{i_r}$. Thus $E = \{e_{i_1}, \ldots, e_{i_r}\}$. The probability of the occurrence of E is denoted by

$P(E)$ and is defined as follows:

$$P(E) = p(e_{i_1}) + \cdots + p(e_{i_r}), \quad \text{or} \quad P(E) = \sum_{a=1}^{r} P(e_{i_a})$$

If E contains only one element, say e_j, it is written as

$$E = \{e_j\} \quad \text{and} \quad P(E) = p(e_j)$$

It is evident, probabilities of events in a finite sample space S are values of an *additive set function* $P(E)$ defined on sets E in S, satisfying the following conditions:

1. If E is any event in S, then
$$P(E) \geq 0 \tag{3.3.2a}$$
2. If E is the sample space S itself, then
$$P(E) = P(S) = 1 \tag{3.3.2b}$$
3. If E and F are two disjoint events in S, then
$$P(E \cup F) = P(E) + P(F) \tag{3.3.2c}$$

These conditions are also sometimes known as *axioms of probability.* In the case of an infinite sample space S, condition 3 extends as follows:
if $E_1, E_2, \ldots$ is an infinite sequence of disjoint events, then

$$P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots \tag{3.3.2d}$$

As E and $\bar{E}$ are disjoint events, then from condition 3, we obtain

$$P(E \cup \bar{E}) = P(E) + P(\bar{E}) \tag{3.3.3}$$

But since $E \cup \bar{E} = S$ and $P(S) = 1$, we have the following:

Theorem 3.3.1 (Rule of complementation) *If* E *is an event in a sample space* S, *then*

$$P(\bar{E}) = 1 - P(E) \tag{3.3.4}$$

The law of complementation provides a simple method of finding the probability of an event $\bar{E}$, if E is an event whose probability is easy to find. We sometimes say that the *odds* in favor of E are

$$\text{Odds}(E) = P(E)/P(\bar{E}) \tag{3.3.4a}$$

which from (3.3.4) takes the form $P(E)/[1-P(E)]$. The reader may note that $0 \leq \text{Odds}(E) \leq \infty$.

Example 3.3.1 (Tossing coins) *Suppose that 10 coins are tossed and we ask for the probability of getting at least 1 head. In this example, the sample space* S *has* $2^{10} = 1024$ *sample points. If the coins are unbiased, the sample points are* equally likely *(sample points are called* equally likely *if each sample point has the same probability of occurring), so that to each of the sample points the probability* 1/1024 *is assigned. If we denote by* E *the event of getting no heads, then* E *contains only one sample point, and* $\bar{E}$*, of course, has 1023 sample points. Thus*

$$P(\bar{E}) = 1 - 1/1024 = 1023/1024$$

The odds on E and $\bar{E}$ are clearly $\text{Odds}(E) = 1/1023$ and $\text{Odds}(\bar{E}) = 1023/1$.
Referring to the statement in Theorem 3.3.1 that $\cup_{i=1}^{k} E_i$ and $\cap_{i=1}^{k} \bar{E}_i$ are disjoint events whose union is S, we have the following rule.

Theorem 3.3.2 (General rule of complementation) *If* $E_1, E_2, \ldots, E_k$ *are events in a* sample space S*, then we have*

$$P\left(\overset{k}{\underset{i=1}{\cap}} \bar{E}_i\right) = 1 - P\left(\overset{k}{\underset{i=1}{\cup}} E_i\right) \tag{3.3.5}$$

Another useful result follows readily from (3.3.2c) by mathematical induction

Theorem 3.3.3 (Rule of addition of probabilities for mutually exclusive events) *If* $E_1, E_2, \ldots, E_k$ *are disjoint events in a sample space* S*, then*

$$P(E_1 \cup E_2 \cup \cdots \cup E_k) = P(E_1) + P(E_2) + \cdots + P(E_k) \tag{3.3.6}$$

Example 3.3.2 (Determination of probabilities of some events) *Suppose that a nickel and a dime are tossed, with* H *and* T *denoting head and tail for the nickel and* h *and* t *denoting head and tail for the dime. The sample space* S *consists of the four elements* Hh, Ht, Th, and Tt. *If these four elements are all assigned equal probabilities and if* E *is the event of getting exactly one head, then* $E = \{Ht\} \cup \{Th\}$*, and we have that*

$$P(E) = P(\{Ht\} \cup \{Th\}) = P(\{Ht\}) + P(\{Th\}) = 1/4 + 1/4 = 1/2$$

Now suppose that E_1 and E_2 are arbitrary events in S. Then from Figure 3.2.2, with $E = E_1$ and $F = E_2$, it can be easily seen that $E_1 \cap E_2, E_1 \cap \bar{E}_2, \bar{E}_1 \cap E_2$, are three disjoint events whose union is $E_1 \cup E_2$. That is,

$$P(E_1 \cup E_2) = P(E_1 \cap E_2) + P(E_1 \cap \bar{E}_2) + P(\bar{E}_1 \cap E_2) \tag{3.3.7}$$

Also, $E_1 \cap \bar{E}_2$ and $E_1 \cap E_2$ are disjoint sets whose union is E_1. Hence,

$$P(E_1) = P(E_1 \cap \bar{E}_2) + P(E_1 \cap E_2) \tag{3.3.8}$$

Similarly

$$P(E_2) = P(\bar{E}_1 \cap E_2) + P(E_1 \cap E_2) \tag{3.3.9}$$

Solving (3.3.8) for $P(E_1 \cap \bar{E}_2)$ and (3.3.9) for $P(\bar{E}_1 \cap E_2)$ and substituting in (3.3.7), we obtain the following.

Theorem 3.3.4 (Rule for addition of probabilities for two arbitrary events) *If E_1 and E_2 are any two events in a sample space* S, *then*

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{3.3.10}$$

The rule for three events $E_1, E_2, and E_3$ is given by

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) \\ - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3) \end{aligned} \tag{3.3.11}$$

More generally, for n events $E_1, \ldots, E_n$, we have,

$$\begin{aligned} P(E_1 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{j>i=1}^{n} P(E_i \cap E_j) \\ + \sum_{k>j>i=1}^{n} P(E_i \cap E_j \cap E_k) + \cdots + (-1)^{n-1} P(E_1 \cap \cdots \cap E_n) \end{aligned} \tag{3.3.12}$$

Note that for $n = 2$, if E_1 and E_2 are disjoint, $P(E_1 \cap E_2) = 0$ and (3.3.10) reduces to (3.3.6); that is,

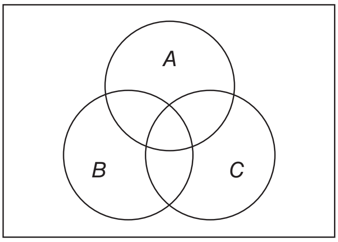
$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Similarly, if E_1, E_2, and E_3 are disjoint, (3.3.11) reduces to

$$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3)$$

PRACTICE PROBLEMS FOR SECTIONS 3.2 AND 3.3

1. Consider a sample space S. Let A and B be any two events in S. Write the expressions in terms of unions, interactions, and complements for the following events:
   (a) At least one of the events A or B occurs.
   (b) Both events A and B occur.
   (c) Neither A nor B occurs.
   (d) Exactly one of the events A or B occurs.
   (e) At most one of the events A or B occurs.
2. Draw a Venn diagram for each event described in Problem 1 above.
3. Describe the sample space for each of the following experiments:
   (a) Three coins are tossed.
   (b) One die is rolled and a coin is tossed.
   (c) Two dice are rolled.
   (d) A family has three children of different ages, and we are interested in recording the gender of these children such that the oldest child is recorded first.
   (e) One die is rolled and two coins are tossed.
4. Two regular dice are rolled simultaneously. If the numbers showing up are different, what is the probability of getting a total of 10 points?
5. Three students are randomly selected from a freshmen engineering class, and it is observed whether they are majoring in chemical, mechanical, or electrical engineering. Describe the sample space for this experiment. What is the probability that at most one of the three students is an EE major?
6. A box contains a shipment of $n(n > 4)$ computer chips, of which four are defective. Four chips are randomly selected and examined as to whether or not the chips are defective. Describe the sample space for this experiment. What is the probability that
   (a) Exactly one of the four chips is defective?
   (b) All four chips are defective?
   (c) Two chips are defective?
7. Given a sample space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and four events A, B, C, and D in S that are defined as $A = \{1, 3, 4, 7\}$, $B = \{2, 4, 6, 8, 9\}$, $C = \{1, 4, 5, 7\}$, and $D = \{1, 3, 5, 7, 9\}$, describe the following events:
   (a) $A \cap B \cap C$, (b) $(A \cap B) \cup (C \cap D)$, (c) $A \cap (B \cup C \cup D)$, (d) $\bar{A} \cap \bar{B}$, (e) $(\overline{A \cup B \cup C \cup D})$, and (f) $(\bar{A} \cap \bar{B} \cap \bar{C} \cap \bar{D})$
8. Given a sample space $S = \{x|3 < x < 10\}$ and two events A and B in S defined as $A = \{x|4 < x < 7\}$ and $B = \{x|5 < x < 9\}$, describe the following events:
   (a) $\bar{A}$, (b) $\overline{A \cup B}$, (c) $A \cup B$ (d) $A \cap B$
9. Suppose that a person is taken to an ER and that A is the event that he is diagnosed with liver cancer, B is the event that he will need a liver transplant, and C is the event that the hospital will find a matching liver on time. The Venn diagram representing these events and various other regions is shown below. Describe in words the events represented by the following regions:



(a) $A \cap B \cap C$, (b) $A \cap (B \cup C)$, (c) $\bar{A} \cap \bar{B}$ (d) $(\bar{A} \cap \bar{B} \cap \bar{C})$

10. In a random experiment four "lock nuts" are selected and each nut is classified either as defective (D) or nondefective (N). Write the sample space for this random experiment.
11. Five women are selected randomly, and their mammograms are examined. Each mammogram is classified as indicating that the woman has breast cancer (C) or does not have breast cancer (N). Write the sample space for this random experiment.
12. The time a biology major takes to dissect a frog is recorded to the nearest minute. Describe the sample space for this random experiment.
13. Three coins are tossed. Describe the following events:
    (a) At least two heads occur.
    (b) At most one head occurs.
    (c) Exactly two heads occur.
    (d) No head occurs.

    Find the probability for the occurrence of each event.
14. Two dice are rolled and the sum of the points that appear on the uppermost faces of the two dice is noted. Write all possible outcomes such that:
    (a) The sum is seven.
    (b) The sum is five or less.
    (c) The sum is even or nine.

    Find the probability for the occurrence of each event you described in parts (a) through (c).

3.4 TECHNIQUES OF COUNTING SAMPLE POINTS

The problem of computing probabilities of events in finite sample spaces where equal probabilities are assigned to the elements reduces to the operation of counting the elements that make up the events in the given sample space. Counting such elements is often greatly simplified by the use of a tree diagram and the rules for permutations and combinations.

3.4.1 Tree Diagram

A tree diagram is a tool that is useful not only in describing the sample points but also in listing them in a systematic way. The following example illustrates this technique.

Example 3.4.1 (Constructing a tree diagram) *Consider a random experiment consisting of three trials. The first trial is testing a chip taken from the production line, the second is randomly selecting a part from the box containing parts produced by six different manufacturers, and the third is, again, testing a chip off the production line. The interest in this experiment is in describing and listing the sample points in the sample space of the experiment.*

Solution: A tree-diagram technique describes and lists the sample points in the sample space of the experiment consisting of three trials. The first trial in this experiment has two possible outcomes: the chip could be defective (D) or nondefective (N); the second trial has six possible outcomes because the part could come from manufacturer 1, 2, 3, 4, 5, or 6; and the third, again, has two possible outcomes (D, N). The problem of constructing a tree diagram for a multitrial experiment is sequential in nature: that is, corresponding to each trial, there is a step of drawing branches of the tree. The tree diagram associated with this experiment is shown in Figure 3.4.1.

The number of sample points in a sample space is equal to the number of branches corresponding to the last trial. For instance, in the present example, the number of sample points in the sample space is equal to the number of branches corresponding to the third trial, which is 24 $(2 \times 6 \times 2)$. To list all the sample points, start counting from o along the paths of all possible connecting branches

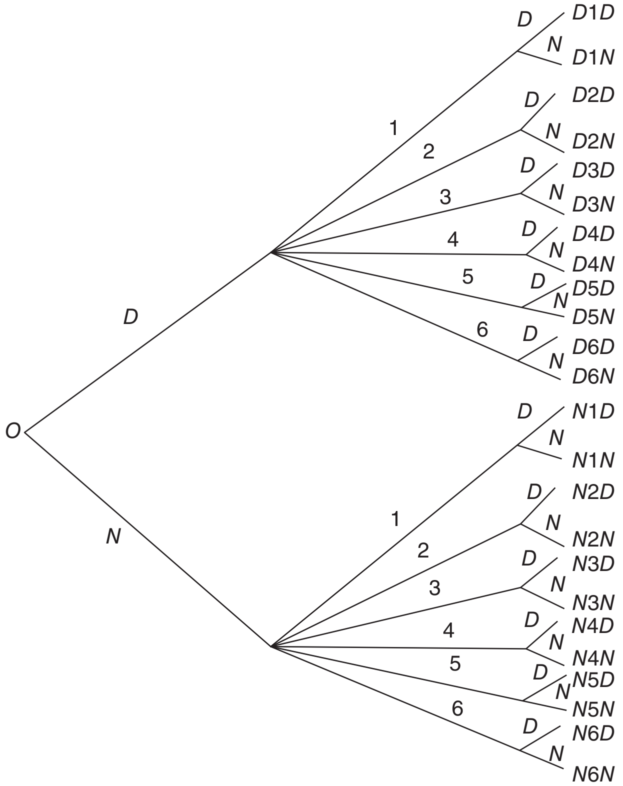


Figure 3.4.1 Tree diagram for the experiment in Example 3.3.1

until the end of the final set of branches, listing the sample points in the same order as the various branches are covered. The sample space S in this example is $S = \{D1D, D1N, D2D, D2N, D3D, D3N, D4D, D4N, D5D, D5N, D6D, D6N, N1D, N1N, N2D, N2N, N3D, N3N, N4D, N4N, N5D, N5N, N6D, N6N\}$.

The tree diagram technique for describing the number of sample points is extendable to an experiment with a large number of trials, where each trial has several possible outcomes. For example, if an experiment has n trials and the ith trial has m_i possible outcomes $(i = 1, 2, 3, \ldots, n)$, then there will be m_1 branches at the starting point o, m_2 branches at the end of each of the m_1 branches, m_3 branches at the end of the each of $m_1 \times m_2$ branches, and so on. The total number of branches at the end would be $m_1 \times m_2 \times m_3 \times \cdots \times m_n$, which represents all the sample points in the sample space S of the experiment. This rule of describing the total number of sample points is known as the *Multiplication Rule.*

3.4.2 Permutations

Suppose that we have n distinct objects $O_1, O_2, \ldots, O_n$. We can determine how many different sequences of x objects can be formed by choosing x objects in succession from the n objects where $1 \leq x \leq n$. For convenience, we may think of a sequence of x places that are to be filled with x objects. We have n choices of objects to fill the first place. After the first place is filled, then with $n - 1$ objects left, we have $n - 1$ choices to fill the second place. Each of the n choices for filling the first place can be combined with each of the $n - 1$ choices for filling the second place, thus yielding $n(n - 1)$ ways of filling the first two places. By continuing this argument, we will see that there are $n(n - 1) \cdots (n - x + 1)$ ways of filling the x places by choosing x objects from the set of n objects. Each of these sequences or arrangements of x objects is called a *permutation* of x objects from n. The total number of permutations of x objects from n, denoted by $P(n, x)$, is given by

$$P(n, x) = n(n - 1) \cdots (n - x + 1) \tag{3.4.1}$$

Note that the number of ways of permuting all the n objects is given by

$$P(n, n) = n(n - 1) \cdots (2)(1) = n! \tag{3.4.2}$$

where $n!$ is read as n factorial.

Expressed in terms of factorials, we easily find that

$$\boxed{P(n, x) = \frac{n!}{(n - x)!} \qquad (3.4.3)}$$

3.4.3 Combinations

It is easy to see that if we select any set of x objects from n, there are $x!$ ways this particular set of x objects can be permuted. In other words, there are $x!$ permutations that contain any set of x objects taken from the n objects. Any set of x objects from n distinct objects is called a combination of x objects from n objects. The number of such combinations is usually denoted by $\binom{n}{x}$. As each combination of each x objects can be permuted in $x!$ ways, these $\binom{n}{x}$ combinations give rise to $\binom{n}{x}x!$ permutations. But this is the total number of permutations when using x objects from the n objects. Hence, $\binom{n}{x}x! = P(n,x)$, so that

$$\boxed{\binom{n}{x} = \frac{P(n,x)}{x!} = \frac{n!}{x!(n-x)!} \qquad (3.4.4)}$$

Example 3.4.2 (Applying concept of combinations) *The number of different possible hands of 13 cards in a pack of 52 ordinary playing cards is the number of combinations of 13 cards from 52 cards, and from (3.4.4) is*

$$\binom{52}{13} = \frac{52!}{13!39!} = 635{,}013,559{,}600$$

Example 3.4.3 (Applying concept of combinations) *The number of samples of 10 objects that can be selected from a lot of 100 objects is*

$$\binom{100}{10} = \frac{100!}{10!90!} = 17{,}310,309{,}456,400$$

Example 3.4.4 (Determining number of combinations) *Suppose that we have a collection of* n *letters in which* x *are A's and $n-x$ are B's. The number of distinguishable arrangements of these* n *letters (*x *A's and $n-x$ B's) written in* n *places is $\binom{n}{x}$.*

We can think of all n places filled with B's, and then select x of these places and replace the B's in them by A's. The number of such selections is $\binom{n}{x}$. This is equivalent to the number of ways we can arrange x A's and $n-x$ B's in n places.

The number $\binom{n}{x}$ is usually called *binomial coefficient*, since it appears in the binomial expansion (for integer $n \geq 1$)

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} \qquad (3.4.5)$$

Example 3.4.5 *The coefficient of $a^x b^{n-x}$ in the expansion of $(a+b)^n$ is $\binom{n}{x}$, since we can write $(a+b)^n$ as*

$$(a+b)(a+b)\cdots(a+b) \quad (n \quad \text{factors}) \qquad (3.4.6)$$

The coefficient of $a^x b^{n-x}$ is the number of ways to pick x of these factors and then choose a from each factor, while taking b from the remaining $(n-x)$ factors.

3.4.4 Arrangements of *n* Objects Involving Several Kinds of Objects

Suppose that a collection of n objects are such that there are x_1 A_1's, x_2 A_2's, $\ldots$, x_k A_k's, where $x_1 + x_2 + \cdots + x_k = n$. Then total number of distinguishable arrangements of these several kinds of A's denoted by $\binom{n}{x_1, \ldots, x_k}$ is

$$\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1! \cdots x_k!} \tag{3.4.7}$$

For if we think of each of the n places being originally filled with objects of type A, there are $\binom{n}{x_1}$ ways of choosing x_1 A's to be replaced by A_1's. In each of these $\binom{n}{x_1}$ ways, there are $\binom{n-x_1}{x_2}$ ways of choosing x_2 A's to be replaced by A_2's. Hence, the number of ways of choosing x_1 A's and replacing them with A_1's and choosing x_2 from the remaining $n - x_1$ A's and replacing them with A_2's is $\binom{n}{x_1}\binom{n-x_1}{x_2}$. Continuing this argument and using equation (3.4.4) shows that the number of ways of choosing x_1 A's and replacing them with A_1's, x_2 A's and replacing them with A_2's, and so on until the last x_k A's replaced with A_k's, is

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-1}}{x_k} = \frac{n!}{x_1!x_2!\cdots x_k!} = \binom{n}{x_1, \ldots, x_k}$$

To illustrate the application of combinations to probability problems involving finite sample spaces, we consider the following example.

Example 3.4.6 (Combinations and probability) *If 13 cards are dealt from a thoroughly shuffled deck of 52 ordinary playing cards, the probability of getting five spades is*

$$\frac{\binom{13}{5}\binom{39}{8}}{\binom{52}{13}}$$

Solution: This result holds because the number of ways of getting five spades from the 13 spades in the deck is $\binom{13}{5}$, and the number of ways of getting 8 nonspades from the 39 nonspades in the deck is $\binom{39}{8}$, and hence, the number of ways five spades and eight nonspades occurs in a hand of 13 cards is the product $\binom{13}{5}\binom{39}{8}$. This is the number of elements in the sample space constituting the event of "getting five spades in dealing 13 cards from a shuffled deck." Since the sample space consists of $\binom{52}{13}$ equally likely sample points, each sample point is assigned the same probability $1/\binom{52}{13}$. Hence, the probability of getting five spades in dealing 13 cards is

$$\frac{\binom{13}{5}\binom{39}{8}}{\binom{52}{13}}$$

PRACTICE PROBLEMS FOR SECTION 3.4

1. In a certain clinical trial, a medical team wants to study four different doses of a new medication for cervical cancer in five patients. In how many different ways can the team select one dose of the medication and one of the patients?

2. A small motel with nine rooms has three twin beds in two rooms, two twin beds in three rooms, and one twin bed in rest of the four rooms. In how many different ways can the manager of the motel assign these rooms to a group of 16 guests who told the manager that they have no preference about whom they share the room with?
3. In how many ways can a class of 30 students select a committee from the class that consists of a president, a vice president, a treasurer, and a secretary (a) if any student may serve either of these roles but no student may serve in multiple roles and (b) if any student may serve in multiple roles?
4. A multiple-choice board exam consists of 15 questions, each question having four possible answers. In how many different ways can a candidate select one answer to each question?
5. A chain restaurant offers a dinner menu with four different soups, three different salads, 10 entrees, and four desserts. In how many ways can a customer choose a soup, a salad, an entrée, and a dessert?
6. If in Problem 3 above, the committee consists of just four members, then in how many ways can the class select the committee?
7. If 13 cards are dealt from a thoroughly shuffled deck of 52 ordinary playing cards, find the probability of getting five spades and four diamonds.
8. How many different permutations can be obtained by arranging the letters of the word engineering? Cardiologist?
9. A cholesterol-lowering drug is manufactured by four different pharmaceutical companies in five different strengths and two different forms (tablet and capsule). In how many different ways can a physician prescribe this drug to a patient?
10. How many different car plates can be issued if the Department of Motor Vehicles decides to first use two letters of the English alphabet and then any four of the digits $0, 1, 2, \ldots, 9$?
11. In a random experiment, one die is rolled, one coin is tossed, and a card is drawn from a well-shuffled regular deck of playing cards and its suit noted. How many sample points are there in the sample space of this random experiment?
12. Each of 10 websites either contains (C) an ad of a car manufacturer or does not contain the ad (N). How many sample points are there in the sample space of a random experiment that selects a website at random?

3.5 CONDITIONAL PROBABILITY

In some probability problems, we are asked to find the probability that an event F occurs when it is known or given that an event E has occurred. This probability, denoted by $P(F|E)$ and called the *conditional probability* of F given E, is obtained essentially by letting E be a new sample space, sometimes known as an induced sample space, and then computing the fraction of probability on E that lies on $E \cap F$, that is,

$$P(F|E) = \frac{P(E \cap F)}{P(E)} \tag{3.5.1}$$

where, of course, $P(E) \neq 0$. Similarly

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{3.5.2}$$

where $P(F) \neq 0$. The conditional probability of $\bar{F}$, given that E has occurred, using (3.5.1), is

$$P(\bar{F}|E) = \frac{P(E \cap \bar{F})}{P(E)}$$

Since we are in the "new" sample space E, we find that (see Figure 3.6.1)

$$P(F|E) + P(\bar{F}|E) = \frac{P(E \cap F) + P(E \cap \bar{F})}{P(E)} = \frac{P(E)}{P(E)} = 1$$

That is, the rule of complementation is preserved in the induced sample space E.

Example 3.5.1 (Concept of conditional probability) *The manufacturing department of a company hires technicians who are college graduates as well as technicians who are not college graduates. Under their diversity program, the manager of any given department is careful to hire both male and female technicians. The data in Table 3.5.1 show a classification of all technicians in a selected department by qualification and gender. Suppose that the manager promotes one of the technicians to a supervisory position.*

(a) If the promoted technician is a woman, then what is the probability that she is a nongraduate?
(b) Find the probability that the promoted technician is a nongraduate when it is not known that the promoted technician is a woman.

Solution: Let S be the sample space associated with this problem, and let E and F be the two events defined as follows:

E: the promoted technician is a nongraduate
F: the promoted technician is a woman

In Part (a) we are interested in finding the conditional probability $P(E|F)$.

Since any of the 100 technicians could be promoted, the sample space S consists of 100 equally likely sample points. The sample points that are favorable to the event E are 65, and those that are favorable to the event F are 44. Also the sample points favorable

Table 3.5.1 Classification of technicians by qualification and gender.

	Graduates	Nongraduates	Total
Male	20	36	56
Female	15	29	44
Total	35	65	100

to both the events E and F are all the women who are nongraduates and equal to 29. To describe this situation, we have

$$P(E) = 65/100, \quad P(F) = 44/100, \quad P(E \cap F) = 29/100$$

(a) Therefore,

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{29/100}{44/100} = 29/44 = 0.659$$

and for part (b), we have that $P(E) = 65/100 = 0.65$.

Note that the probability $P(E)$, sometimes known as the *absolute probability* of E, is different from the conditional probability *P(E|F)*. If the conditional probability *P(E|F)* is the same as the absolute probability *P(E)*, that is, *P(E|F) = P(E)*, then the two events E and F are said to be *independent*. In this example, the events E and F are not independent.

Definition 3.5.1 Let S be a sample space, and let E and F be any two events in S. The events E and F are called *independent* if and only if any one of the following is true:

$$1.\ P(E|F) = P(E) \tag{3.5.3}$$

$$2.\ P(F|E) = P(F) \tag{3.5.4}$$

$$3.\ P(E \cap F) = P(E) \times P(F) \tag{3.5.5}$$

The conditions in equations (3.5.3)–(3.5.5) are equivalent in the sense that if one is true, then the other two are true. We now have the following theorem, which gives rise to the so-called *multiplication rule*.

Theorem 3.5.1 (Rule of multiplication of probabilities) *If* E *and* F *are events in a sample space* S *such that* $P(E) \neq 0$, *then*

$$P(E \cap F) = P(E) \cdot P(F|E) \tag{3.5.6}$$

Equation (3.5.6) provides a two-step rule for determining the probability of the occurrence of $(E \cap F)$ by first determining the probability of E and then multiplying by the conditional probability of F given E.

Example 3.5.2 (Applying probability in testing quality) *Two of the light bulbs in a box of six have broken filaments. If the bulbs are tested at random, one at a time, what is the probability that the second defective bulb is found when the third bulb is tested?*

Solution: Let E be the event of getting one good and one defective bulb in the first two bulbs tested, and let F be the event of getting a defective bulb on drawing the third bulb.

Then, $(E \cap F)$ is the event whose probability we are seeking. Hence,

$$P(E \cap F) = P(E) \cdot P(F|E)$$

The sample space S in which E lies consists of all possible selections of two bulbs out of six, the number of elements in S being $\binom{6}{2} = 15$. The event E consists of all selections of one good and one defective bulb out of four good and two defective bulbs, and the number of such selections is $4 \times 2 = 8$. Therefore, $P(E) = 8/15$. We now compute $P(F|E)$, the probability that F occurs given that E occurs. If E has occurred, there are three good and one defective bulb left in the box. F is the event of drawing the defective one on the next draw, that is, from a box of four that has three good bulbs and one defective. Thus, $P(F|E) = 1/4$. The required probability is

$$P(E \cap F) = P(E) \cdot P(F|E) = (8/15) \times (1/4) = 2/15$$

For the case of three events, $E_1, E_2, and\, E_3$, the extended form of (3.5.6) is

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_2|E_1) \cdot P(E_3|E_1 \cap E_2) \tag{3.5.7}$$

provided that $P(E_1)$ and $P(E_1 \cap E_2)$ are both not zero. Formula (3.5.7) extends to any finite number of events $E_1, \ldots, E_k$. Note that equation (3.5.5) also can be extended to the case of several mutually independent events $E_1, \ldots, E_n$ so that for these n independent events

$$P(E_1 \cap \cdots \cap E_n) = P(E_1) \cdots P(E_n) \tag{3.5.8}$$

Example 3.5.3 (Rolling a die n times) *If a true die is thrown* n *times, what is the probability of never getting an ace (one-spot)?*

Solution: Let E_1 be the event of not getting an ace on the first throw, E_2 the event of not getting an ace on the second throw, and so on. Assuming independence of the events $E_1, \ldots, E_n$ and a "true" die, we have $P(E_1) = \cdots = P(E_n) = 5/6$. Hence, the required probability from (3.5.8) is

$$P(E_1 \cap \cdots \cap E_n) = P(E_1) \cdots P(E_n) = (5/6)^n$$

3.6 BAYES'S THEOREM

An interesting version of the conditional probability formula (3.5.1) comes from the work of the Reverend Thomas Bayes. Bayes's result was published posthumously in 1763.

Suppose that E and F are two events in a sample space S and such that $E \cap F \neq \emptyset$. From the Venn diagram in Figure 3.6.1, we can see that the events $(E \cap F)$ and $(E \cap \bar{F})$ are disjoint and that their union is E, so that

$$P(E) = P(E \cap F) + P(E \cap \bar{F}) \tag{3.6.1}$$

Using the rule given by (3.5.6), we can rewrite equation (3.6.1) as

$$P(E) = P(F)P(E|F) + P(\bar{F})P(E|\bar{F}) \tag{3.6.2}$$

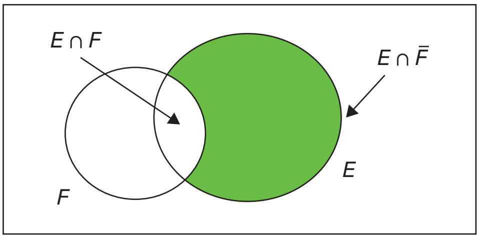


Figure 3.6.1 Venn diagram showing events $(E \cap F)$ and $(E \cap \overline{F})$.

We can rewrite (3.5.1) in the form

$$P(F|E) = \frac{P(F)P(E|F)}{P(F)P(E|F) + P(\bar{F})P(E|\bar{F})} \tag{3.6.3}$$

The rule provided by (3.6.3) is known as Bayes's theorem for two events E and F; the probabilities $P(F)$ and $P(\bar{F})$ are sometimes referred to as the *prior* probabilities of events F and $\bar{F}$, respectively (note that $F \cup \bar{F} = S, F \cap \bar{F} = \emptyset$). The conditional probability $P(F|E)$ as given by Bayes's theorem (3.6.3), is referred to as the *posterior* probability of F, given that the event E has occurred. An interpretation of (3.6.3) is that, *posterior* to observing that the event E has occurred, the probability of F changes from $P(F)$, the prior probability, to $P(F|E)$, the posterior probability.

Example 3.6.1 (Bayes's theorem in action) *The Gimmick TV model A uses a printed circuit, and the company has a routine method for diagnosing defects in the circuitry when a set fails. Over the years, the experience with this routine diagnostic method yields the following pertinent information: the probability that a set that fails due to printed circuit defects (PCD) is correctly diagnosed as failing because of PCD is 80%. The probability that a set that fails due to causes other than PCD has been diagnosed incorrectly as failing because of PCD is 30%. Experience with printed circuits further shows that about 25% of all model A failures are due to PCD. Find the probability that the model A set's failure is due to PCD, given that it has been diagnosed as being due to PCD.*

Solution: To answer this question, we use Bayes's theorem (3.6.3) to find the posterior probability of a set's failure being due to PCD, after observing that the failure is diagnosed as being due to a faulty PCD. We let

$F =$ event, set fails due to PCD
$E =$ event, set failure is diagnosed as being due to PCD

and we wish to determine the posterior probability $P(F|E)$.

We are given that $P(F) = 0.25$ so that $P(\bar{F}) = 0.75$, and that $P(E|F) = 0.80$ and $P(E|\bar{F}) = 0.30$. Applying (3.6.3) gives

$$\begin{aligned} P(F|E) &= \frac{P(F)P(E|F)}{P(F)P(E|F) + P(\bar{F})P(E|\bar{F})} \\ &= \frac{(0.25)(0.80)}{(0.25)(0.80) + (0.75)(0.30)} = 0.471 \end{aligned}$$

Notice that in light of the event E having occurred, the probability of F has changed from the *prior* probability of 25% to the *posterior* probability of 47.1%.

Formula (3.6.3) can be generalized to more complicated situations. Indeed Bayes stated his theorem for the more general situation, which appears below.

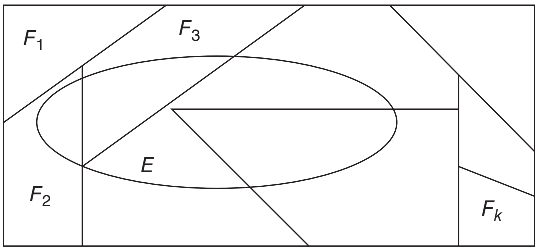


Figure 3.6.2 Venn diagram showing $F_1, F_2, \ldots, F_k$ mutually exclusive events in S.

Theorem 3.6.1 (Bayes's theorem) *Suppose that* $F_1, F_2, \ldots, F_k$ *are mutually exclusive events in* S *such that* $\sum_{i=1}^{k} P(F_i) = 1$, *and* E *is any other event in* S. *Then*

$$P(F_i|E) = \frac{P(F_i)P(E|F_i)}{\sum_{i=1}^{k} P(F_i)P(E|F_i)} \tag{3.6.4}$$

We note that (3.6.3) is a special case of (3.6.4), with $k = 2$, $F_1 = F$, and $F_2 = \bar{F}$. Bayes's theorem for k events F_i has aroused much controversy. The reason for this is that in many situations, the *prior* probabilities $P(F_i)$ are unknown. In practice, when not much is known about a priori, these have often been set equal to $1/k$ as advocated by Bayes himself. The setting of $P(F_i) = 1/k$ in what is called the "in-ignorance" situation is the source of the controversy. Of course, when the $P(F_i)$'s are known or may be estimated on the basis of considerable past experience, (3.6.4) provides a way of incorporating prior knowledge about the F_i to determine the conditional probabilities $P(F_i|E)$ as given by (3.6.4). We illustrate (3.6.4) with the following example.

Example 3.6.2 (Applying Bayes's theorem) *David, Kevin, and Anita are three doctors in a clinic. Dr. David sees 40% of the patients, Dr. Anita sees 25% of the patients, and 35% of the patients are seen by Dr. Kevin. Further 10% of Dr. David's patients are on Medicare, while 15% of Dr. Anita's and 20% of Dr. Kevin's patients are on Medicare. It is found that a randomly selected patient is a Medicare patient. Find the probability that he/she is Dr. Kevin's patient.*

Solution: Let

F_1 = Person is Dr. Kevin's patient
E = Person is a Medicare patient

and let

F_2 = Person is Dr. Anita's patient
F_3 = Person is Dr. David's patient

We are given that $P(F_1) = 0.35$, $P(F_2) = 0.25$, $P(F_3) = 0.40$ while $P(E|F_1) = 0.20$, $P(E|F_2) = 0.15$ and $P(E|F_3) = 0.10$. We wish to find $P(F_1|E)$. Using (3.6.4) with $k = 3$, we have

$$P(F_1|E) = \frac{P(F_1)P(E|F_1)}{\sum_{i=1}^{3} P(F_i)P(E|F_i)} = \frac{(0.35)(0.20)}{(0.35)(0.20) + (0.25)(0.15) + (0.40)(0.10)} = 0.475$$

We note that the posterior probability of F_1, given E, is 0.475, while the prior probability of F_1 was $P(F_1) = 0.35$. We sometimes say that the prior information about F_1 has been updated in light of the information that E occurred to the posterior probability of F_1, given E, through Bayes's theorem.

PRACTICE PROBLEMS FOR SECTIONS 3.5 AND 3.6

1. A regular die is rolled. If the number that showed up is odd, what is the probability that it is 3 or 5?
2. Three balanced coins are tossed simultaneously. What is the probability that exactly two heads appear given that at least one head has appeared?
3. Suppose that A_1, A_2, A_3, A_4, and A_5 are five mutually exclusive and exhaustive events in a sample space S, and suppose that $P(A_1) = 0.2, P(A_2) = 0.1, P(A_3) = 0.15, P(A_4) = 0.3$, and $P(A_5) = 0.25$. Another event E in S is such that $P(E|A_1) = 0.2, P(E|A_2) = 0.1, P(E|A_3) = 0.35, P(E|A_4) = 0.3$, and $P(E|A_5) = 0.25$. Find the probabilities $P(A_1|E), P(A_2|E), P(A_3|E), P(A_4|E)$, and $P(A_5|E)$.
4. Suppose that four attorneys A_1, A_2, A_3, and A_4 deal with all the criminal cases in a district court. The following table gives the percentages of the cases that each of these attorneys handles, and also the probability that each loses the case;

Attorney	Probability of handling the case	Probability of losing the case
A_1	0.40	0.15
A_2	0.25	0.30
A_3	0.25	0.20
A_4	0.10	0.40

   Suppose that a criminal case was lost in the court. Find the probability that this case was handled by Attorney A_2.

5. Suppose that a random experiment consists of randomly selecting one of four coins C_1, C_2, C_3, and C_4, tossing it and observing whether a head or a tail occurs. Further suppose that the coins C_1, C_2, and C_3 are biased such that the probabilities of a head occurring for coins C_1, C_2, and C_3 are 0.9, 0.75, and 0.60, respectively, while the fourth coin C_4 is a fair coin.

(a) If the outcome of the experiment was a head, find the probability that coin C_2 was tossed.
(b) If the outcome of the experiment was a tail, find the probability that coin C_4 was tossed.

6. An industry uses three methods, M_1, M_2, and M_3 to manufacture a part. Of all the parts manufactured, 45% are produced by method M_1, 32% by method M_2, and the rest 23% by method M_3. Further it has been noted that 3% of the parts manufactured by method M_1 are defective, while 2% manufactured by method M_2 and 1.5% by method M_3 are defective. A randomly selected part is found to be defective. Find the probability that the part was manufactured by (a) method M_1, (b) method M_2.
7. There are four roads connecting location A and location B. The probabilities that if a person takes Road I, Road II, Road III, or Road IV from location A to B, then he/she will arrive late because of getting stuck in the traffic are 0.3, 0.20, 0.60, and 0.35, respectively. Suppose that a person chooses a road randomly and he/she arrives late. What is the probability that the person chose to take Road III?
8. Suppose that in a ball-bearing manufacturing plant four machines M_1, M_2, M_3, and M_4 manufacture 36%, 25%, 23%, and 16% of the ball bearings, respectively. It is observed that the four machines produce 2%, 2.5%, 2.6%, and 3% defective ball bearings, respectively. If the ball bearings manufactured by these machines are mixed in a well-mixed lot and then a randomly selected ball bearing is found to be defective, find the probability that the defective ball bearing is manufactured by (a) machine M_1, (b) machine M_2, (c) machine M_3, (d) machine M_4.
9. An urn contains five coins of which three are fair, one is two-headed and one is two-tailed. A coin is drawn at random and tossed twice. If a head appears both times, what is the probability that the coin is two-headed?

3.7 INTRODUCING RANDOM VARIABLES

Suppose that a finite sample space S consists of m elements $e_1, e_2, \ldots, e_m$. There are 2^m possible events that can be formed from these elements, provided that the empty event $\emptyset$ and the entire sample space S are counted as two of the events. This is revealed by the fact that we have the choice of selecting or not selecting each of the m elements in making up an event. Rarely, if ever, is one interested in all these 2^m events and their probabilities. Rather, the interest lies in a relatively small number of events produced by specified values of some function defined over the elements of a sample space. For instance, in the sample space S of the $\binom{52}{13}$ possible hands of 13 bridge cards, we are usually interested in events such as getting two aces, or eight spades, or 10 honor cards, and so on.

A real and single-valued function $X(e)$ defined on each element e in the sample space S is called a *random variable*. Suppose that $X(e)$ can take on the values $x_1, x_2, \ldots, x_k$. Let $E_1, E_2, \ldots, E_k$ be the events that are mutually exclusive and exhaustive in the sample space S, for which $X(e) = x_1, X(e) = x_2, \ldots, X(e) = x_k$, respectively. Let $P(E_1) = p(x_1), \ldots, P(E_k) = p(x_k)$. Then, we say that $X(e)$ is a random variable defined over the sample space S and is a *discrete random* variable which takes the values $x_1, x_2, \ldots, x_k$ with the probabilities $p(x_1), p(x_2), \dots, p(x_k)$, respectively. Since $E_1, E_2, \dots, E_k$ are disjoint and their union is equal to the entire sample space S, we have

$$p(x_1) + \cdots + p(x_k) = 1 \tag{3.7.1}$$

We can arrange the values of X and the corresponding probabilities in table form as follows:

$X = x$	x_1	x_2	$\cdots$	x_k
$p(x)$	$p(x_1)$	$p(x_2)$	$\cdots$	$p(x_k)$

(3.7.2)

The values of the discrete random variable $X(e)$ together with their associated probabilities are called the discrete distribution of $X(e)$. The function $p(x_i)$, defined by

$$P(X(e) = x_i) = p(x_i), \quad i = 1, 2, \dots, k \tag{3.7.3}$$

is called the *probability function* (p.f.) of $X(e)$. Ordinarily, we drop the e and refer to the random variable as X. The set of possible values $x_1, x_2, \dots, x_k$ is called the *sample space of the random variable X.*

Example 3.7.1 (Defining concept of the probability function) *Let* X *be a random variable denoting the sum of the number of dots that appear when two dice are thrown. If each of the 36 elements in the sample space is assigned the same probability, namely* $1/36$*, then* $p(x)$*, the probability function of* X*, is as follows:*

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

If the sample space S has an infinite number of elements and if the random variable x can take on a countably infinite set of values $x_1, x_2, \dots,$ we have a *discrete* random variable with sample space –$x_1, x_2, \dots$″.

Example 3.7.2 (Probability function for an event to occur) *Let* X *be a random variable denoting the number of times a die is thrown until an ace appears. The sample space of* X *is* $1, 2, \dots,$ *and the probability function* $p(x) = P(X = x)$ *is given by the table:*

x	1	2	$\cdots$	x	$\cdots$
$p(x)$	$1/6$	$5/6^2$	$\cdots$	$5^{x-1}/6^x$	$\cdots$

since

$$p(x) = \frac{5^{x-1}}{6^{x-1}} \times \frac{1}{6} = \frac{5^{x-1}}{6^x}, \quad x = 1, 2, \dots$$

Note that the probability function $p(x)$ must possess the following properties.

$$\text{(i)} \quad p(x) \geq 0 \quad \text{and} \quad \text{(ii)} \quad \sum_x p(x) = 1$$

If any one of these properties does not hold, then $p(x)$ is not a probability function.

We conclude this section with the comment that in this section, we have discussed only *discrete random variables.* There is indeed another type of random variables, called *continuous random variables*, discussed extensively in Chapter 5. Suffice it to say here that a continuous random variable may take all values in at least one interval, and it, of course, contains an infinite number of values that are not countable. This is in contrast with a discrete random variable, which takes values that are countable, as discussed here and in Chapter 4.

Review Practice Problems

1. Certain pieces made by an automatic lathe are subject to three kinds of defects X, Y, Z. A sample of 1000 pieces was inspected with the following results: 2.1% had type X defect, 2.4% had type Y defect, and 2.8% had type Z defect. 0.3% had both type X and type Y defects, 0.4% had both type X and type Z defects, and 0.6% had both type Y and type Z defects. 0.1% had type X, type Y, and type Z defects.
   Then find:
   (a) What percent had none of these defects?
   (b) What percent had at least one of these defects?
   (c) What percent were free of type X and/or type Y defects?
   (d) What percent had not more than one of these defects?

2. Two inspectors A and B independently inspected the same lot of items. Four percent of the items are actually defective. The results turn out to be as follows: 5% of the items are called defective by A, and 6% of the items are called defective by B. 2% of the items are correctly called defective by A, and 3% of the items are correctly called defective by B. 4% of the items are called defective by both A and B, and 1% of the items are correctly called defective by both A and B.
   (a) Make a Venn diagram showing percentages of items in the eight possible disjoint classes generated by the classification of the two inspectors and the true classification of the items.
   (b) What percent of the truly defective items are missed by inspectors?

3. (a) A box of 100 items contains 90 nondefective items, seven with type A defects, five with type B defects, and two with both types of defects. Let S be the sample space generated by the operation of drawing one item blindly from the box. Let E_A be the event of getting a type A defective and E_B the event of getting a type B defective. Set up a Venn diagram and show the following events: $E_A \cap E_B, E_A \cap \bar{E}_B, E_A \cap \bar{E}_B, and\, \bar{E}_A \cap \bar{E}_B$ indicating the number of elements in each. How many elements are in $E_A \cup E_B$?
   (b) In the box described in part (a) of this problem, let S^* be the sample space generated by blindly drawing two items simultaneously from the box. Let G be the event that corresponds to pairs of nondefective items. Let G_A be the event that corresponds to pairs of items containing at least one type A defective while

G_B is the event that corresponds to pairs of items containing at least one type B defective. Set up a Venn diagram and show the eight basic events in S^* obtained by placing or omitting a bar over each label in $G \cap G_A \cap G_B$, and write in the number of elements in each of the eight basic events.

4. A "true" icosahedral (Japanese) die has 20 sides, two sides marked with 0, two sides with 1, two sides with 2, ..., two sides with 9. The probabilities assigned to the 20 faces are all equal. Suppose then that three such dice are thrown. Find the following probabilities:
   (a) That no two top faces will be alike.
   (b) That at least two top faces will be alike.
   (c) That all three top faces will be different even numbers (0 is considered an even number).
   (d) Generalize parts (a) and (b) to the case of a true $2n$-sided die with two sides marked 1, two sides marked 2, ..., two sides marked n.

5. Ten defective items are known to be in a box of 100 items.
   (a) If they are located by testing the items one at a time until all defectives are found, what is the probability that the 10th (last) defective item is located when the 50th item is tested?
   (b) What is the probability that if 50 items are drawn at random from the box and tested, all 10 defectives will be found?
   (c) If 20 are tested and found to be nondefective, what is the probability that all defectives will be found among the next 30 tested?

6. If a lot of 1000 articles has 100 defectives and if a sample of 10 articles is selected at random from the lot, what is the probability that the sample will contain:
   (a) No defectives?
   (b) At least one defective?

7. Assume that a given type of aircraft motor will operate eight hours without failure with probability 0.99. Assume that a two-motor plane can fly with at least one motor, that a four-motor plane can fly with at least two motors, and that failure of one motor is independent of the failure of another.
   (a) If a two-motor plane and a four-motor plane take off for an eight hour flight, show that the two-motor plane is more than 25 times more likely to be forced down by motor failure than the four-motor plane.
   (b) Compute the respective probabilities that the planes will not be forced down by motor failure.
   (c) What is the answer to (b) if the probability of failure of a motor during an 8-hour period is p rather than 0.01?

8. Suppose 10 chips are marked $1, 2, \ldots, 10$, respectively, and put in a hat. If two chips are simultaneously drawn at random, what is the probability that
   (a) Their difference will be exactly 1?
   (b) Neither number will exceed 5?
   (c) Both numbers will be even?
   (d) At least one of the numbers will be 1 or 10?

9. If the probability is 0.001 that a type-X 20-W bulb will fail in a 10-hour test, what is the probability that a sign constructed from 1000 such bulbs will burn 10 hours:

(a) With no bulb failures?
(b) With one bulb failure?
(c) With k bulb failures?

10. The game of craps is played with two ordinary six-sided dice as follows: If the shooter throws 7 or 11, he wins without further throwing; if he throws 2, 3, or 12, he loses without further throwing. If he throws 4, 5, 6, 8, 9, or 10, he must continue throwing until a 7 or the "point" he initially threw appears. If after continuing a 7 appears first he loses; if the "point" he initially threw appears first, he wins. Show that the probability is approximately 0.4929 that the shooter wins (assuming true dice).

11. In a group of 11 persons, no two persons are of the same age. We are to choose five people at random from this group of 11.
    (a) What is the probability that the oldest and the youngest persons of the 11 will be among those chosen?
    (b) What is the probability that the third youngest of the 5 chosen will be the 6th youngest of the 11?
    (c) What is the probability that at least three of four of the youngest of the 11 will be chosen?

12. Suppose that the probability is 1/365 that a person selected at random was born on any specified day of the year (ignoring persons born on February 29). What is the probability that if r people are randomly selected, no two will have the same birthday? (The smallest value of r for which the probability that at least two will have a common birthday exceeds 0.5 is 23.)

13. Suppose that six true dice are rolled simultaneously. What is the probability of getting
    (a) All faces alike?
    (b) No two faces alike?
    (c) Only five different faces?

14. A, B, C, and D are four events that are such that $P(A) = P(B) = P(C) = P(D) = p_1$, $P(A \cap B) = P(A \cap C) = \cdots = P(C \cap D) = p_2$ $P(A \cap B \cap C) = P(A \cap B \cap D) = P(A \cap C \cap D) = P(B \cap C \cap D) = p_3$, $P(A \cap B \cap C \cap D) = p_4$. Express the values of the following probabilities in terms of $p_1, p_2, p_3, and\, p_4$:
    (a) $P(A \cup B \cup C)$
    (b) $P(A \cup B \cup C \cup D)$
    (c) $P(A \cap B | C \cap D)$
    (d) Probability of the occurrence of exactly 1, exactly 2, exactly 3, of the events A, B, C, D.

15. If four addressed letters are inserted into four addressed envelopes at random, what is the probability that
    (a) No letter is inserted into its own envelope?
    (b) At least one letter is inserted into its own envelope?

16. Three machines A, B, and C produce 40%, 45%, and 15%, respectively, of the total number of nuts produced by a certain factory. The percentages of defective output of these machines are 3%, 6%, and 9%. If a nut is selected at random, find the probability that the item is defective.

17. In Problem 16, suppose that a nut is selected at random and is found to be defective. Find the probability that the item was produced by machine A.

18. Enrollment data at a certain college shows that 30% of the men and 10% of the women are studying statistics and that the men form 45% of the student body. If a student is selected at random and is found to be studying statistics, determine the probability that the student is a woman.

19. A certain cancer diagnostic test is 95% accurate on those that do have cancer, and 90% accurate on those that do not have cancer. If 0.5% of the population actually does have cancer, compute the probability that a particular individual has cancer if the test finds that he has cancer.

20. An urn contains three blue and seven white chips. A chip is selected at random. If the color of the chip selected is white, it is replaced and two more white chips are added to the urn. However, if the chip drawn is blue, it is not replaced and no additional chips are put in the urn. A chip is then drawn from the urn a second time. What is the probability that it is white?

21. Referring to Problem 20, suppose that we are given that the chip selected for the second time is white. What is the probability that the chip selected at the first stage is blue?

22. Referring to Problem 5, suppose that it takes 11 tests to find all the 10 defectives, that is, the 11th test produces the last defective. What is the probability that the first item is nondefective?

23. A bag contains a nickel, quarter, and "dime", with the dime being a fake coin and having two heads. A coin is chosen at random from the bag and tossed four times in succession. If the result is four heads, what is the probability that the fake dime has been used?

24. In a playground, there are 18 players, 11 of them boys and seven are girls. Eight of the boys and three of the girls are soccer players; the rest are basketball players. The name of each player is written on a separate slip, and then these slips are put into an urn. One slip is drawn randomly from the urn; the player whose name appears on this slip is given a prize. What is the probability that a soccer player gets the prize given that a boy gets the prize?

25. Let A_1, A_2, A_3, and A_4 be mutually exclusive and exhaustive events in a sample space S, and let $P(A_1) = 0.2, P(A_2) = 0.1, P(A_3) = 0.4$, and $P(A_4) = 0.3$. Let B be another event in S such that $P(B|A_1) = 0.4, P(B|A_2) = 0.1, P(B|A_3) = 0.6, and\, P(B|A_4) = 0.2$. Find the probabilities $P(A_1|B), P(A_2|B), P(A_3|B)$, and $P(A_4|B)$.

26. An industry uses three methods M_1, M_2, and M_3 to train their workers. Of all the workers trained, 50% are trained by method M_1, 28% by method M_2, and the rest, 22%, by method M_3. Further 10% of those trained by method M_1 do not perform their job well, while 5% trained by method M_2 and 15% by methods M_3 also do not perform their job well. A randomly selected worker does not perform his job well. Find the probability that the worker was trained by (a) method M_1, (b) method M_2, and (c) method M_3.

27. Suppose that an insurance company finds that during the past three years, 60% of their policy holders have no accident, 25% had one accident, and the remaining 15%

had two or more accidents. Further, from their past 10 year history, the insurance company concludes that those who did not have any accident have a 1% chance of having an accident in the future, while those who had one accident have a 3% chance of having an accident in the future. Those who had two or more accidents in the past three years have 10% chances having an accident in the future. Suppose that shortly after the above study, one of their policy holders has an accident. What is the probability that this policy holder did not have any accident in the past three years?

28. Suppose that in a bolt manufacturing plant, machines *A, B*, and *C* manufacture, respectively, 45%, 25%, and 30% of the bolts. It is observed that the three machines produce, respectively, 5%, 2%, and 3% defective bolts. If the bolts produced by these machines are mixed in a well-mixed lot and then a randomly selected bolt is found to be defective, find the probability that the defective bolt is produced by (a) machine *A*, (b) machine *B*, and (c) machine *C*.

29. A poll was conducted among 1000 registered voters in a metropolitan area asking their position on bringing a casino in that city. The results of the poll are shown in the following table:

Sex	Percentage of voters polled	Percentage favoring casino	Percentage not favoring casino	Percentage having no opinion
Male	55%	75%	20%	5%
Female	45%	40%	50%	10%

What is the probability that a registered voter selected at random
(a) Was a man, given that the voter favored the casino?
(b) Was a woman, given that the voter has no opinion about the casino?
(c) Was a woman, given that the voter did not favor the bringing in of a casino?

30. Let a random variable be distributed as shown below.

$X = x$	0	1	2	3	4	5	6
$p(x)$	0.1	0.09	0.2	0.15	0.16	0.2	

(a) Find the probability $p(6)$.
(b) Find the probability $P(3 \leq X \leq 5)$.
(c) Find the probability $P(X \leq 4)$.
(d) Find the probability $P(X > 2)$.

31. Determine which of the following distributions do not represent a probability distribution. Justify your answer.
(a)

$X = x$	0	1	2	3	4	5
$p(x)$	0.1	0.09	0.2	0.15	0.16	0.2

(b)

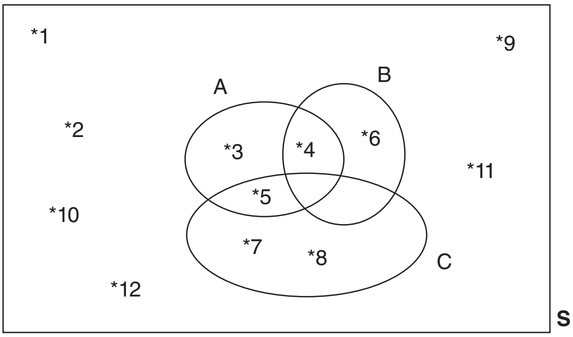
$X = x$	0	1	2	3	4	5
$p(x)$	-0.1	0.09	0.3	0.15	0.16	0.4

(c)

$X = x$	0	1	2	3	4	5
$p(x)$	0.2	0.09	0.2	0.15	0.16	0.2

32. In Problem 31, construct a bar chart for the probability distribution that satisfies the conditions of a probability distribution.

33. The following Venn diagram describes the sample space S of a random experiment and events A, B, and C associated with the experiment. (S consists of 12 elements, denoted by $e_i = {}^*i, \quad i = 1, 2, \ldots, 12.$)



(a) Express the events A, B, and C in terms of the elements ${}^*i$.
(b) Suppose that the sample points in the sample space S are equally likely.

Find the following probabilities:

(i) $P(A)$
(ii) $P(B)$
(iii) $P(C)$
(iv) $P(A \cap B)$
(v) $P(A \cap B \cap C)$
(vi) $P(\overline{A \cap B \cap C})$
(vii) $P(A \cup B \cup C)$
(viii) $P(\overline{A \cup B \cup C})$

Chapter 4

DISCRETE RANDOM VARIABLES AND SOME IMPORTANT DISCRETE PROBABILITY DISTRIBUTIONS

The focus of this chapter is a discussion of some important discrete probability distributions.

Topics Covered

- Discrete random variables and some important probability distributions
- Approximation of the binomial by the Poisson distribution
- Determination of the cumulative distribution functions (c.d.f.) from probability functions
- Determination of the mean and variance of different discrete random variables
- Determination of the probabilities of events involving discrete random variables using the statistical packages MINITAB, R, and JMP

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand various important discrete distributions and apply them to determine probabilities in real-world problems.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

- Determine approximate probabilities of rare events.
- Determine the mean and the variance of discrete random variables using general techniques and moment-generating functions.
- Apply the statistical packages MINITAB, R, and JMP to calculate probabilities when using different discrete probability models.

4.1 GRAPHICAL DESCRIPTIONS OF DISCRETE DISTRIBUTIONS

In Chapter 2, we discussed methods of describing *empirical distributions*, that is, distributions of the numerical values of the measurements obtained in a sample. In Chapter 3, we discussed basic concepts of probability theory. In this chapter, we discuss methods that can be used for describing *theoretical discrete distributions.* In Section 3.7, we introduced the concept of a discrete random variable. The set of possible values of a discrete random variable X, together with their associated probabilities (see Section 3.7), is called the *probability function* (p.f.) of the discrete random variable. Discrete distributions are conveniently described graphically. Thus, in general, suppose that the probability function of a discrete random variable X is as described below in (4.1.1):

$X = x$	x_1	x_2	$\cdots$	x_k
$p(x)$	$p(x_1)$	$p(x_2)$	$\cdots$	$p(x_k)$

(4.1.1)

The probability graph for this discrete probability function is shown in Figure 4.1.1, where the lengths of the vertical lines represent the magnitudes of the probabilities.

The *cumulative distribution function* (c.d.f.), $F(x)$, provides an alternative way of describing a discrete random variable X. For any real number x, $F(x)$ is defined as follows:

$$F(x) = P(X \leq x) = \sum_{i} p(x_i) \tag{4.1.2}$$

where $\sum_{i}$ denotes summation for all values of i for which $x_i \leq x$.

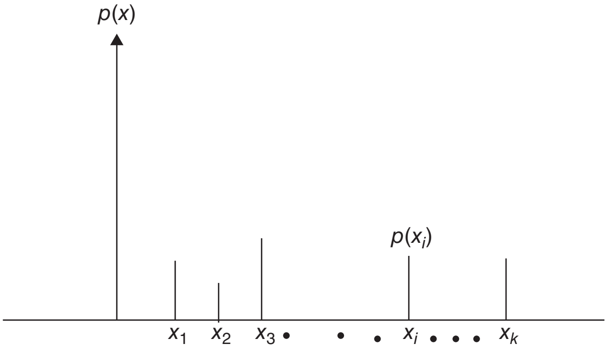


Figure 4.1.1 Graphical representation of the discrete probability function in (4.1.1).

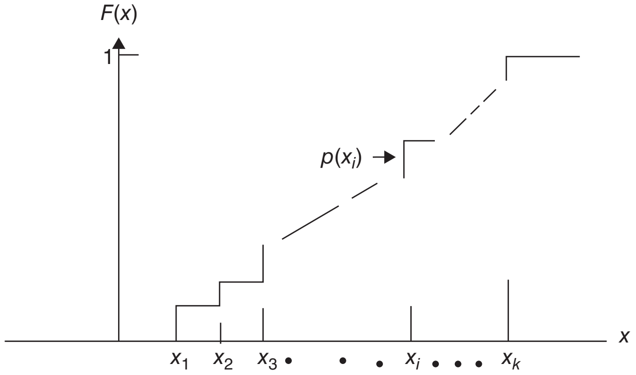


Figure 4.1.2 Graph of cumulative distribution function $F(x)$ of a discrete random variable X.

It will be observed that

$$\begin{aligned} F(x_1) &= p(x_1) \\ F(x_2) &= p(x_1) + p(x_2) \\ &\vdots \\ F(x_i) &= p(x_1) + \cdots + p(x_i) \qquad (4.1.3) \\ &\vdots \\ F(x_k) &= p(x_1) + \cdots + p(x_k) = 1 \end{aligned}$$

It can easily be seen that $F(x)$ is a step function defined for *all* (real) values of x having a graph as shown in Figure 4.1.2. Note that the steps in the graph occur at the x points for which $x = x_1, x_2, \ldots, x_k$.

Actually the c.d.f. is a more convenient device for describing the probability distributions of what are called *continuous random variables* (discussed in Chapter 5). But the introduction of $F(x)$ for the case of discrete random variables at this point simplifies the interpretation of $F(x)$ for a continuous random variable (see Chapter 5).

4.2 MEAN AND VARIANCE OF A DISCRETE RANDOM VARIABLE

4.2.1 Expected Value of Discrete Random Variables and Their Functions

In Chapter 2, we defined the average of a sample of measurements. In a similar way, we define the *population mean* $E(X)$ (denoted by μ) of the discrete random variable X having probability function given in (4.1.1) as follows:

$$\mu = E(X) = \sum_{i=1}^{k} x_i p(x_i) \tag{4.2.1}$$

Clearly, $E(X)$ defined in (4.2.1) is a weighted average of the possible values $x_1, x_2, \ldots, x_k$ of the random variable X, where the weights are the associated probabilities $p(x_1), p(x_2), \ldots, p(x_k)$. $E(X)$ is often called the expected value or expectation of the random variable X.

Note that the word *average* identifies $\bar{X} = \sum_{i=1}^{n} X_i/n$, a quantity that *depends on the sample alone*. We call such quantities *statistics*. Furthermore, in this book, the word *mean* is used to identify a characteristic of the population (4.1.1) from which the sample is to be drawn, where the $p(x_i)$ gives the theoretical distribution of the population values. We call theoretical quantities such as μ the *parameters* of the distribution.

Corresponding to the procedure for defining the variance S^2 of a sample of measurements, we define the *variance* of the discrete random variable X, $Var(X)$ (denoted by σ^2), as follows:

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - \mu)^2 p(x_i) \tag{4.2.2a}$$

$$\sigma^2 = \sum_{i=1}^{k} x_i^2 p(x_i) - \mu^2 = E(X^2) - \mu^2 \tag{4.2.2b}$$

It should be noted that $Var(X)$ is a weighted average of the squared deviations of $x_1, x_2, \ldots, x_k$ from the mean μ, the weights again being the probabilities. A useful alternative expression for σ^2 is given by (4.2.2b). This alternative expression is obtained by expanding the squared term $(x_i - \mu)^2$ in (4.2.2a), performing the summation, and simplifying the results.

So far, we have defined the mean and variance of discrete random variables. More generally, if X is a discrete random variable, as defined in Section 4.1, and $g(X)$ is a function (real and single-valued) of X, the mean value or expectation of $g(X)$, written as $E(g(X))$, is defined by

$$E(g(X)) = \sum_{i=1}^{k} g(x_i) p(x_i) \tag{4.2.3}$$

If we take, for example, $g(X) = X^r$, then (4.2.3) takes the form

$$E(X^r) = \sum_{i=1}^{k} x_i^r p(x_i) = \mu_r' \tag{4.2.4}$$

The quantity μ'_r defined by (4.2.4) is called the rth moment of the random variable X about the origin. Note that if $r = 0$, we have

$$E(X^0) = E(1) = 1$$

and if $r = 1$, we have

$$E(X) = \mu$$

Often, $E(X^r)$ is denoted by μ'_r, as mentioned in (4.2.4).

Now, if we put $g(X) = (X - \mu)^r$, (4.2.3) takes the form

$$E[(X-\mu)^r] = \sum_{i=1}^{k} (x_i - \mu)^r p(x_i) \tag{4.2.5}$$

The quantity defined by (4.2.5) is called the rth moment of X about its mean and denoted by μ_r. Note that $r = 0$ gives $E[(X - \mu)^0] = 1$, and if $r = 1$, it is easily proved that $\mu_1 = 0$. Also note that $\mu_2 = E[(X - \mu)^2]$ is the variance of X which we denote by σ^2. The reader should verify that if c is a constant, then (see Theorem 4.2.2)

$$E(cX) = cE(X), \quad Var(cX) = c^2 Var(X) \tag{4.2.6}$$

We now provide some general results in the form of theorems about the expected values of discrete random variables.

Theorem 4.2.1 *Let c be a constant and X be a discrete random variable distributed with probability function $p(x)$. Then,*

$$E(c) = \sum cp(x) = c \sum p(x) = c \times 1 = c \tag{4.2.7}$$

Theorem 4.2.2 *Let* c *be a constant and $g(X)$ be a function of a discrete random variable* X *that is distributed with probability function $p(x)$. Then,*

$$E(cg(X)) = cE(g(X)) \tag{4.2.8}$$

Theorem 4.2.3 *Let $g_i(X), i = 1, 2, \ldots, n$ be* n *functions of a discrete random variable* X *that is distributed with probability function $p(x)$. Then,*

$$E\left[\sum_{i=1}^{n} g_i(X)\right] = \sum_{i=1}^{n} E[g_i(X)] \tag{4.2.9}$$

Proof:

$$E\left[\sum_{i=1}^{n} g_i(X)\right] = \sum_x \left[\sum_{i=1}^{n} g_i(x)\right] \times p(x) = \sum_{i=1}^{n}\left[\sum_x g_i(x) \times p(x)\right] = \sum_{i=1}^{n} E[g_i(X)]$$

□

4.2.2 The Moment-Generating Function-Expected Value of a Special Function of X

Referring to (4.2.3), suppose that we set $g(X) = e^{Xt}$ and find its expected value; then

$$E(e^{Xt}) = \sum_{i=1}^{n} e^{x_i t} p(x_i) \tag{4.2.10}$$

The function of t so obtained is called the *moment-generating function* of the random variable X, and is denoted by $M_X(t)$, that is,

$$M_X(t) = E(e^{Xt}) \tag{4.2.11}$$

Note that $M_X(t)$ can be written as

$$\begin{aligned} M_X(t) &= E\left(1 + Xt + \frac{X^2t^2}{2!} + \cdots + \frac{X^kt^k}{k!} + \cdots\right) \\ &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \cdots + \frac{t^k}{k!}E(X^k) + \cdots \end{aligned}$$

Thus, we have

$$M_X(t) = 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \cdots + \frac{t^k}{k!}\mu'_k + \cdots \tag{4.2.12}$$

if the moments $\mu'_1, \mu'_2, \ldots,$ are all finite, then the coefficient of $t^k/k!$ in the expansion of the moment generating function (m.g.f.) is the kth moment about the origin of X. If we differentiate $M_X(t)$ k times, we obtain (assuming differentiability under summation signs)

$$\frac{d^k}{dt^k}M_X(t) = E(X^k e^{Xt}) \tag{4.2.13}$$

If we then set $t = 0$, we have

$$\frac{d^k}{dt^k}M_X(0) = E(X^k) = \mu'_k \tag{4.2.14}$$

We now mention some important properties of $M_X(t)$ the m.g.f. of X. First, if we are interested in the random variable cX, where c is any constant, then by definition, the m.g.f. of cX is

$$M_{cX}(t) = E(e^{cXt}) = E(e^{Xct}) = M_X(ct) \tag{4.2.15}$$

Second, the m.g.f. for $X + a$, where a is a constant, is

$$M_{X+a}(t) = E(e^{(X+a)t}) = E(e^{at}e^{Xt}) = e^{at}M_X(t) \tag{4.2.16}$$

Note that (4.2.16) enables us to find moments of X about its mean by setting $a = -\mu$.

PRACTICE PROBLEM FOR SECTIONS 4.1 AND 4.2

1. A hospital is known for coronary artery bypass grafting. Let X be the number of such surgeries done on a given day. The following table gives the probability distribution of the random variable X:

$X = x$	0	1	2	3	4	5
$p(x)$	0.02	0.05	0.10	0.15	0.18	0.50

   Find the following probabilities:
   (a) $P(X \leq 2)$
   (b) $P(2 < X < 5)$
   (c) $P(X \geq 2)$
   (d) $P(1 \leq X \leq 4)$

2. Each of the following tables lists the values of a random variable X and presumably their corresponding probabilities. Determine whether or not each represents a probability distribution. Justify your answer.

   (a)

$X = x$	2	3	4	5
$p(x)$	0.15	0.25	0.35	0.45

   (b)

$X = x$	4	7	8	9
$p(x)$	0.15	0.15	0.45	0.30

(c)

$X = x$	2	5	7	9	11
$p(x)$	0.10	0.15	0.35	0.23	0.17

3. Each of the following tables lists the values of a random variable X and presumably their corresponding probabilities. Determine whether or not each represents a probability distribution. Justify your answer.

(a)

$X = x$	1	2	3	4	5
$p(x)$	0.12	0.17	0.31	0.23	0.17

(b)

$X = x$	0	1	2	4
$p(x)$	0.15	0.15	0.33	0.47

(c)

$X = x$	2	3	4	5
$p(x)$	?0.05	0.25	0.35	0.45

4. A manufacturer of car parts ships five exhaust gas temperature sensors, each of which is independently declared to be either conforming or not conforming. Assume that the probability that a sensor is conforming is 0.75. Let X be the number of sensors of the five shipped that are conforming. Find the probability distribution of the random variable X.
5. Refer to Problem 4. Find the mean and the variance of the random variable X.
6. Suppose that the moment-generating function of a random variable X is given by

$$M_X(t) = (0.4e^t + 0.6)^{10}$$

Determine the mean and variance of the random variable X.

7. A pair of fair dice is rolled and a random variable X, the sum of points that turn up, is observed. Find the probability distribution of the random variable X and determine the mean and the variance of the distribution you obtained.
8. A company developed a new shampoo to reduce dandruff. Six persons tried that shampoo. Assume that the probability that a person gets some relief is 0.60. Let X be the number of persons out of the six who find some relief. Determine the probability distribution of the random variable X and then find its mean and variance.
9. Refer to Problem 7. Find the mean and the variance of the random variable (a) $Y = 3X$ and (b) $Y = 2X + 5$.
10. Suppose that the moment-generating function of a random variable X is given by

$$M_X(t) = (0.3e^t + 0.7)^{10}$$

Find the moment-generating function of a random variable (a) $Y = 2X$ and (b) $Y = 3X + 5$.

4.3 THE DISCRETE UNIFORM DISTRIBUTION

The discrete uniform distribution is perhaps the simplest discrete probability distribution. Consider a discrete random variable X and let $x_1, x_2, \ldots, x_i, \ldots, x_N$ be the values that it can assume. Then X is said to be distributed by the uniform distribution if it assumes each of its possible values x_i, $i = 1, 2, 3, \ldots, N$ with equal probability.

Here, the probability function of a discrete uniform distribution is given below in (4.3.1), and its graphical representation is shown in Figure 4.3.1.

$X = x$	x_1	x_2	$\cdots$	x_N
$p(x)$	$1/N$	$1/N$	$\cdots$	$1/N$

(4.3.1)

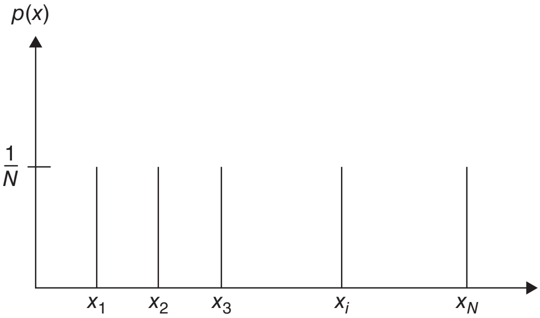


Figure 4.3.1 Graph of uniform probability function given in (4.3.1).

Example 4.3.1 (Uniform distribution) *Consider an experiment of tossing a fair die and observing the number that appears. The sample space of this experiment is* $S = \{1, 2, 3, 4, 5, 6\}$*, and each element of the sample space occurs with probability 1/6. Thus, in this example, the random variable* X*, denoting the number that appears on the die, is distributed by the uniform distribution, which is written as*

$X = x$	1	2	$\cdots$	6
$p(x)$	1/6	1/6	$\cdots$	1/6

The mean and variance of the discrete uniform distribution with probability function given in equation (4.3.1) are given by

$$\mu = E(X) = \sum_{i=1}^{N} x_i p(x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma^2 = E(X-\mu)^2 = \sum_{i=1}^{N} (x_i - \mu)^2 p(x_i)$$
$$= \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

respectively. In general, if $x_i = 1, 2, 3, \ldots, N$, then

$$\boxed{\mu = (N+1)/2 \quad \text{and} \quad \sigma^2 = (N^2-1)/12 \qquad (4.3.2)}$$

Example 4.3.2 (Mean and variance of the uniform distribution) *The mean and variance of the random variable X in Example 4.3.1 are obtained by using (4.3.2).*

That is,

$$\mu = (N+1)/2 = (6+1)/2 = 3.5 \quad \text{and} \quad \sigma^2 = (N^2-1)/12 = (6^2-1)/12 = 2.917$$

4.4 THE HYPERGEOMETRIC DISTRIBUTION

The hypergeometric probability function provides probabilities of certain events when a sample of n objects is drawn at random from a finite population of N objects, where the sampling is conducted *without replacement*, and where each element of the population may be dichotomized in some simple fashion as belonging to one of two disjoint classes.

As an example of sampling from a dichotomous population, we consider first a hand of bridge and ask for the probability of getting exactly x spades. Here, the dichotomy is "spades" and "nonspades (i.e., clubs, hearts, and diamonds)"; the population consists of $N = 52$ cards; the sample consists of $n = 13$ cards. Or, consider a lot of N transistors, of which $100p\%$ are defective and $100(1-p)\%$ are nondefective, where $0 \le p \le 1$. If these are packaged in boxes of n transistors, interest may be in finding the probability that x transistors in a given box are defective. Here, the population is the lot of N transistors, and the sample size is n. The dichotomy in this example is, of course, "defective" and "nondefective."

In general, suppose that a random sample of n observations is taken from a population of size N without replacement and that the observations are examined as to having an attribute A or not having this attribute. Objects of the former kind will be called type A and the latter type $\bar{A}$. Suppose that N_1 objects of the population are of type A and $N_2 = N - N_1$ are of type $\bar{A}$. We want to determine the probability that x of the objects in a random sample of size n are of type A (and thus $n - x$ are of type $\bar{A}$).

We know that the total number of ways of choosing a random sample (without replacement) of n from N is $\binom{N}{n}$, which is the number of elements in the sample space S for this problem. Each of these elements will be assigned the probability $1/\binom{N}{n}$. The number of ways of choosing x objects of type A from N_1 is $\binom{N_1}{x}$, $0 \le x \le N_1$, and the number of ways of obtaining $n-x$ objects of type $\bar{A}$ from N_2 is $\binom{N_2}{n-x}$, $0 \le n-x \le N_2 = N - N_1$. Each of these latter ways may be combined with each of the former so that the number of elements in S comprising the event of getting x objects of type A is $\binom{N_1}{x}\binom{N-N_1}{n-x}$. Note that x is a value of a random variable X. Hence, the probability we seek is given by

$$h(x) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}} \tag{4.4.1}$$

The sample space of X is the set of integers x that satisfies the inequalities $\max[0, n-(N-N_1)] \le x \le \min(n, N_1)$, since $0 \le x \le N_1$ and $0 \le n - x \le N - N_1$.

Expression (4.4.1) is the probability function of the *hypergeometric distribution.* It can easily be seen that the sum of $h(x)$ over all values of x in the sample space of X is 1. Further we note that $h(x)$ is the probability of obtaining x objects of type A in a random sample of size n, when sampling *without* replacement from a population of size N that contains N_1 objects of type A and $N_2 = N - N_1$ objects of type $\bar{A}$.

Example 4.4.1 (Applying concept of hypergeometric distribution) *Let us consider the probability of getting x spades in a hand of bridge. In the framework above,* A *denotes a spade, and $\bar{A}$ denotes any card other than a spade. Hence, $N = 52, N_1 = 13, N_2 = N - N_1 = 39$. If* X *denotes the number of spades in a hand of $n = 13$ cards,* X *has the sample space* $\max[0, 13 - (52 - 13)] \leq x \leq \min[13, 13]$*, that is, $0 \leq x \leq 13$, and X has probability function given by*

$$h(x) = \frac{\binom{13}{x}\binom{39}{13-x}}{\binom{52}{13}}$$

For instance, if $x = 0$, then the probability of getting 0 spades in a hand is given by

$$h(0) = \frac{\binom{13}{0}\binom{39}{13}}{\binom{52}{13}} = 0.0128$$

In engineering applications, it is more convenient to consider the hypergeometric distribution from the following point of view. Consider a dichotomized finite population or lot of size N, consisting of Np objects of type A and $N(1-p)$ objects of type $\bar{A}$, where $0 \leq p \leq 1$. Note that N_1 and $N_2 = N - N_1$ have been replaced by Np and $N(1-p)$, respectively, where $p = N_1/N$. If a sample of size n is drawn without replacement, then the probability of obtaining x objects of type A is

$$h(x) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}} \tag{4.4.2}$$

The sample space of X consists of the integers x that satisfy the inequalities $\max[0, n - N(1-p)] \leq x \leq \min(n, Np)$.

The *mean* and the *variance* of the hypergeometric distribution with probability function in equation (4.4.2) are given by

$$\mu = np, \quad \sigma^2 = \left(\frac{N-n}{N-1}\right) npq, \quad p = N_1/N, \quad q = 1 - p \tag{4.4.3}$$

The derivations of (4.4.3) appear in Section 4.10.

Example 4.4.2 (Detecting defective light bulbs) *A carton contains 24 light bulbs, 12.5% of which are defective. What is the probability that, if a sample of six is chosen at random from the carton of the bulbs, then $x = 0, 1, 2, 3$ will be defective?*

Solution: Using (4.4.2), we have, since 12.5% of 24 is 3,

$$h(x) = \frac{\binom{3}{x}\binom{21}{6-x}}{\binom{24}{6}}, \quad 0 \le x \le 3$$

This gives

$$h(0) = \frac{\binom{3}{0}\binom{21}{6}}{\binom{24}{6}} = 0.40316, \quad h(1) = \frac{\binom{3}{1}\binom{21}{5}}{\binom{24}{6}} = 0.45356$$

$$h(2) = \frac{\binom{3}{2}\binom{21}{4}}{\binom{24}{6}} = 0.13340, \quad h(3) = \frac{\binom{3}{3}\binom{21}{3}}{\binom{24}{6}} = 0.00988$$

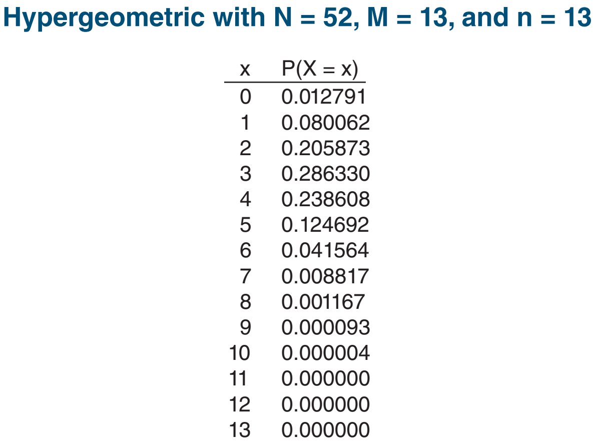
Note here that $\sum_{x=0}^{3} h(x) = 1$.

Example 4.4.3 (Using MINITAB and R) *Refer to Example 4.4.1. Find the probabilities of getting* $x = 0, 1, 2, 3, \ldots, 13$, *spades in a hand of bridge.*

Solution: Again, in the framework of this example, A denotes a spade, and $\bar{A}$ denotes any card other than a spade. Hence, $N_1 = 13, N - N_1 = 39$. Let X denote the number of spades in a hand of $n = 13$ cards. To find the probabilities of obtaining $x = 0, 1, 2, 3, \ldots, 13$ spades in a hand of bridge, proceed as follows:

MINITAB

1. Enter the vales $0, 1, 2, \ldots, 13$ in column C1.
2. From the Menu bar, select **Calc** > **Probability Distributions** > **Hypergeometric**.
3. In the dialog box that appears on the screen, click the circle next to **Probability**.
4. Enter 52 in the box next to **Population size (N)**, 13 in the box next to **Event count in population (M)** (this is the population size of category A), and, again, 13 in the box next to **Sample size (n)**.
5. Click the circle next to **Input column** and type C1 in the box next to it. Click **OK**. The desired probabilities will show up in the Session Window as:



```
Hypergeometric with N = 52, M = 13, and n = 13

 x   P(X = x)
 0   0.012791
 1   0.080062
 2   0.205873
 3   0.286330
 4   0.238608
 5   0.124692
 6   0.041564
 7   0.008817
 8   0.001167
 9   0.000093
10   0.000004
11   0.000000
12   0.000000
13   0.000000
```

USING R

R has a built in ‘dhyper(x, m, n, k)’ function in ‘stats’ library that can be used to calculate the Hypergeometric probabilities, where x is the number of type A objects being selected, m is the number of type A objects in the population, n is the number of type $\bar{A}$ objects in the population, and k is the total number of objects being selected. Probabilities that $x = 0, 1, 2, 3, \ldots, 13$ can be obtained by running the following R code in the R Console window.

```
prob = dhyper(c(0:13), 13, 39, 13)

round(prob, 6)

#R output

 [1]  0.012791  0.080062  0.205873  0.286330  0.238608  0.124692  0.041564
 [8]  0.008817  0.001167  0.000093  0.000004  0.000000  0.000000  0.000000
```

PRACTICE PROBLEMS FOR SECTIONS 4.3 AND 4.4

1. A box contains 10 chips numbered from 1 to 10. A chip is selected (with replacement) randomly and the number on that chip is jotted down. Let this number be denoted by a random variable X. Determine the probability distribution of X and then find its mean and variance.
2. Suppose that the probability distribution of a random variable X is as shown below. Determine the mean and variance of the random variable $Y = 3X$.

$X = x$	1	2	3	4	5	6	7	8
$p(x)$	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

3. Suppose that the random variable X has a hypergeometric distribution with $N = 12, n = 4$, and $N_1 = 6$. Determine the probability distribution of the random variable X and then find its mean and variance.
4. A shipment contains 20 assembled circuit boards of which five are defective. Ten circuit boards from the shipment are selected without replacement. Suppose that X denotes the number of defective boards out of the 10 selected. Find the probability distribution of the random variable X and then find its mean and variance.
5. Twenty identical chips marked as 1, 2, ..., 20 are put in a container and mixed well. A chip is drawn randomly and the number on the chip is observed. Find the following probabilities that the observed number is (a) greater than 15, (b) between 10 and 18 (inclusive), (c) less than 10.
6. Referring to Problem 5, let a random variable X denote the number observed. Find the probability distribution of the random variable X and determine the mean and the variance of the distribution you obtained.
7. A manager of a manufacturing company has 8 female and 12 male engineers in her department. The manager randomly selected a team of six engineers to attend a business meeting. Find the probability that the team had (a) two female engineers and (b) at least three female engineers.

8. An IRS inspector randomly selects five persons from a group of 20 who are potential candidates to be audited. Of the 20 persons, nine were audited in the past, while the other 11 have never been audited before. Find the probability that the number of persons selected who were audited in the past is (a) exactly three, (b) more than two, (c) at least two, (d) at most three.
9. A movie store has five fiction and seven other movies on display. A customer selects four of these 12 at random. What is the probability that the number of fiction movies among the four selected movies is (a) exactly two, (b) between two and four (inclusive), (c) at most two.
10. An electronic company ships a lot of 50 computer hard drives to a store. At the arrival of the shipment, the store manager selects at random three hard drives to test. If the lot had five defective hard drives, find the probability that the number of defective hard drives among the three selected is (a) exactly one, (b) none, (c) at most one.

4.5 THE BERNOULLI DISTRIBUTION

Consider a random experiment E consisting of repeated trials where each trial has only two possible outcomes, referred to as success S and failure F. Then, a sequence of independent trials (repetitions), where the probability of success on each trial remains a constant p and the probability of failure is $(1-p)$, which is called a sequence of Bernoulli trials (note that the probability of failure $(1-p)$ is commonly denoted by q so that $p+q=1$). For example, if we toss a coin repeatedly, we would have Bernoulli trials; in each trial, the probability of a head as well as of a tail remains constant.

Let X be a random variable denoting a success or failure in each Bernoulli trial. Clearly, if we set $X = 1$ or 0, if the trial is observed to be a success or a failure, respectively, then

$$P(X=1)=p, \quad P(X=0)=1-p=q \tag{4.5.1}$$

Thus, the probability function of the Bernoulli random variable X is given by (4.5.1), which may be summarized as shown in (4.5.2).

$$p(x)=p^x q^{1-x}, \quad x=0,1 \tag{4.5.2}$$

A random variable X is said to be distributed by the *Bernoulli* distribution if its probability function is defined as in Equation (4.5.2). Note that the *Bernoulli* distribution is also sometimes known as a point binomial distribution.

The *mean* and *variance* of a Bernoulli distribution are given, respectively, by

$$\mu=p, \quad \sigma^2=pq \tag{4.5.3}$$

It can easily be shown that the moment-generating function of the Bernoulli distribution is

$$M_X(t) = pe^t + q \tag{4.5.4}$$

4.6 THE BINOMIAL DISTRIBUTION

The binomial distribution is one of the most commonly used discrete probability distributions. It is applicable whenever an experiment possesses the following characteristics:

1. The experiment consists of n independent trials.
2. Each trial has two possible outcomes, usually referred to as success and failure.
3. The probability of success, p, for each trial is constant throughout the experiment, and consequently the probability q of failure is constant throughout the experiment.

The probabilities given by the binomial distribution may arise in the following ways:

1. Sampling from a finite population with replacement
2. Sampling from an infinite population (often referred to as an indefinitely large population), with or without replacement.

Suppose that we wish to draw a sample of size n from a lot of N objects of which Np are defectives and Nq are nondefectives, and the sampling is done with replacement. In other words, we draw at random a member of the lot, examine it, record the result, and replace it in the lot, "mix thoroughly," and repeat the procedure a further $n-1$ times. We want to determine the probability of obtaining x defectives in the sample of n trials.

For example, if D denotes the event "defective" and $\bar{D}$ "nondefective," then the sample space consists of the 2^n possible sequences of n outcomes, a D, or a $\bar{D}$. Thus, x defectives would occur in n trials in some sequence of x D's and $(n-x)$ $\bar{D}$'s, such as

$$D\bar{D}DD\bar{D}\bar{D}\cdots D\bar{D}$$

Since the trials are independent, the probability associated with such a sequence is

$$p \times q \times p \times p \times q \times q \times \cdots \times p \times q$$

where, of course, there are x factors having the value p, and $n-x$ factors having the value q. Thus, the probability of such a sequence is

$$p^x q^{(n-x)} \tag{4.6.1}$$

Now, there are $\binom{n}{x}$ different possible sequences of x D's and $(n-x)$ $\bar{D}$'s in the sample space. The probability that any one of these occurs is $p^x q^{(n-x)}$. Since these $\binom{n}{x}$ different sequences are mutually exclusive elements in the sample space, the probability of obtaining x defectives in n trials is therefore given by

$$b(x) = \binom{n}{x} p^x q^{(n-x)}, \quad 0 \leq x \leq n \tag{4.6.2}$$

Now let X be the number of defectives obtained in the n trials. Then, the sample space of X is $0, 1, \ldots, n$, that is X, the number of D's has values ranging over the elements of the sample space and is called a *binomial random variable*, with $P(X = x) = b(x)$.

Thus, the expression given by (4.6.2) is the *probability function* of the binomial random variable X with parameter $0 < p < 1$ and $x = 0, 1, 2, \ldots, n$. It derives its name from the fact that the probabilities are the successive terms in the expansion of the binomial $(q + p)^n$ since the $(x + 1)$st term in the expansion of the (q+p)^n is the expression for $b(x)$. Thus,

$$\sum_{x=0}^{n} b(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{(n-x)} = [q + p]^n = 1$$

that is, the sum of $b(x)$ over all points in the sample space of X is unity.

The *mean* and *variance* of the binomial distribution with probability function given in Equation (4.6.2) are given by

$$\mu = np, \quad \sigma^2 = npq \tag{4.6.3}$$

The derivations of (4.6.3) appear in Section 4.10.

Example 4.6.1 (Gambling and probabilities) *Two dice are thrown 100 times, and the number of nines is recorded. What is the probability that* x *nines occur? That at least three nines occur?*

Solution: It is apparent that we are examining each roll of the two dice for the events nine or non-nine. The probability of obtaining a nine by throwing two dice is $4/36 = 1/9$, that is, $p = 1/9$. Hence, the probability that x nines occur in 100 throws of the two dice is

$$b(x) = \binom{100}{x} \left(\frac{1}{9}\right)^x \left(\frac{8}{9}\right)^{100-x}; \quad x = 0, 1, 2, \ldots, 100$$

In answer to the second question of at least three nines appearing, we have

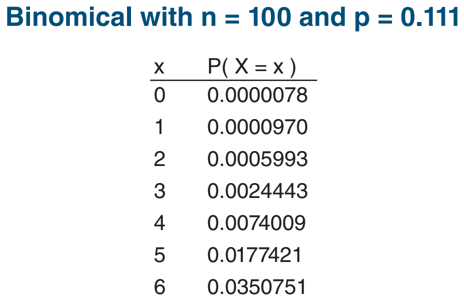
$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) = 1 - P(X \leq 2) \\ &= 1 - \sum_{x=0}^{2} \binom{100}{x} \left(\frac{1}{9}\right)^x \left(\frac{8}{9}\right)^{100-x} \\ &= 1 - (0.000008 + 0.000097 + 0.000599) \\ &= 1 - 0.0007 = 0.9993 \end{aligned}$$

Example 4.6.2 (Applying MINITAB and R) *Refer to Example 4.6.1. Using MINITAB and R, find the probability that* x *number of nines occur, where $x = 0, 1, 2, \ldots, 6$.*

Solution: From our discussion in Example 4.6.1, we have $n = 100$, $p = 1/9 = 0.111$. We want to find the probability $P(X = x)$, where $x = 0, 1, 2, \ldots, 6$. To find these probabilities, we proceed as follows:

MINITAB

1. Enter the values 0, 1, 2, . . . , 6 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distributions** > **Binomial**.
3. In the dialog box, click the circle next to **Probability.**
4. Enter 100 (the number of trials) in the box next to **Number of trials** and 0.111 (the probability of success) in the box next to **Event Probability**.
5. Click the circle next to **Input column** and type C1 in the box next to it.
6. Click **OK**. The desired probabilities will show up in the Session window as:



```
Binomical with n = 100 and p = 0.111

x    P( X = x )
0    0.0000078
1    0.0000970
2    0.0005993
3    0.0024443
4    0.0074009
5    0.0177421
6    0.0350751
```

If we want to store these probabilities in a column, say in column C2, then type C2 in the box next to **Optional storage**.

USING R

R has a built in ‘dbinom(x, size, prob)’ function in ‘stats’ library that can be used to calculate the binomial probabilities, where x is the number of successes, size is the total number of trials, and prob is probability of success in a single trial. Probabilities that $x = 0, 1, 2, 3, \ldots, 6$ can be obtained by running the following R code in the R Console window.

```
prob = dbinom(c(0:6), 100, .111)
round(prob, 6)

#R output

 [1] 0.000008 0.000097 0.000599 0.002444 0.007401 0.017742 0.035075
```

Theorem 4.6.1 *The moment-generating function of the binomial random variable* X *is given by*

$$M_X(t) = (pe^t + q)^n \tag{4.6.4}$$

Proof:

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} = (pe^t + q)^n$$

□

PRACTICE PROBLEMS FOR SECTIONS 4.5 AND 4.6

1. A sample of 16 PCV valves for gas engines is randomly selected from a very large batch and tested. Let X denote the number of valves out of the 16 selected that are found defective when tested. Valves are defective or nondefective independently. If the probability of a valve being defective is 0.02, find the following probabilities:
   (a) No valve is defective.
   (b) At the most one valve is defective.
   (c) At least one valve is defective.
2. Suppose that the probability that a patient admitted in a hospital is diagnosed with a certain type of cancer is 0.03. Suppose that on a given day 10 patients are admitted and X denotes the number of patients diagnosed with this type of cancer. Determine the probability distribution of the random variable X. Find the mean and the variance of X.
3. Define Bernoulli trials in words. Suppose that the probability of success of a Bernoulli trial is p, and you are interested in determining the probability of X successes in n independent Bernoulli trials. Describe the probability distribution of the random variable X.
4. Six missiles are fired at a certain target. The probability that a missile will hit the target is 75%. What is the probability that of the six missiles fired (a) exactly five will hit the target, (b) at least three will hit the target, (c) all six will hit the target?
5. The mean and the variance of a binomial distribution with parameters n and p are 12 and 3. Find the following probabilities: (a) $P(X < 4)$, (b) $P(4 \leq X \leq 11)$, (c) $P(X \geq 7)$.
6. A multiple choice test consists of 12 questions with each question having three choices. If a student checks the answers randomly, find the probability that the student gets (a) five correct answers, (b) between four and eight (inclusive) correct answers, (c) at most five correct answers.
7. In Problem 5, clearly state the probability distribution of the binomial random variable X. What is the probability that the random variable X falls in the interval $[\mu - 3\sigma, \mu + 3\sigma]$?
8. In a survey conducted by a social worker, 30% of the women responded that they were the victims of domestic violence. Assume that this percentage is true for all

women in the United States. Using the binomial probability distribution (Table A.2), find the probability that the number of women in a random sample of 20 who were victims of domestic violence is (a) at least 6, (b) between 7 and 11 (inclusive), (c) at most 8.

9. An electronic system is designed to work as long as at least five of its eight major components function. Each of these components work independently with probability 0.6. What is the probability that the system will work?
10. It has been claimed that at least 40% of all the personal bankruptcies in the United States are due to medical bills. If in a given county, 10 personal bankruptcies occur during a certain period, then find the probability that at least four of these bankruptcies are due to the medical bills.

4.7 THE MULTINOMIAL DISTRIBUTION

Suppose that a trial results in one and only one of k mutually exclusive events $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$, respectively, where $p_1 + p_2 + \cdots + p_k = 1$. If n independent trials are made, then it is seen, by an argument similar to that by which the probability function (4.6.2) of the binomial distribution is obtained, that the probability of getting x_1 $E_1's, x_2$ $E_2's, \ldots, x_k$ $E_k's$ is given by

$$m(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots\ x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{4.7.1}$$

where $0 \le x_i \le n, i = 1, 2, \ldots, k$, and $x_1 + x_2 + \cdots + x_k = n$. This is the probability function of the *multinomial distribution.* The name derives from the fact that the probabilities are the terms in the expansion of $(p_1 + p_2 + \cdots + p_k)^n$. Note that if $k = 2$, we have the binomial distribution, and hence the multinomial distribution is essentially an extension of the binomial distribution.

Example 4.7.1 (Probabilities for a trinomial situation) *Consider the production of ball bearings of a certain type whose diameters should be 0.2500 in. Because of the inherent variability in the manufacturing process and because of consumer demand, the bearings are classified as* undersize, oversize, *or* acceptable *if they measure less than 0.2495 in., more than 0.2505 in., or between 0.2495 and 0.2505 in., respectively. Suppose that the production process for these bearings is such that 4% of the bearings are undersize, 6% are oversize, and 90% are acceptable. If 100 of these bearings are picked at random, the probability of getting x_1 undersize, x_2 oversize, and x_3 acceptable bearings is given by*

$$m(x_1, x_2, x_3) = \frac{100!}{x_1!x_2!x_3!}(0.04)^{x_1}(0.06)^{x_2}(0.90)^{x_3} \tag{4.7.2}$$

where $0 \le x_1 \le 100, 0 \le x_2 \le 100, 0 \le x_3 \le 100,$ *and* $\sum_{i=1}^{3} x_i = 100.$

Note that (4.7.2) is an example of a multivariate probability function; a more thorough discussion of multivariate probability functions is given in Chapter 6.

PRACTICE PROBLEMS FOR SECTION 4.7

1. In a certain city, too many accidents occur every year, so the motor vehicle department has been quite strict about passing or failing the persons who take their driving test. The probabilities that a person who takes this test will pass in the first, second, or third attempt are 0.25, 0.30, and 0.45, respectively. What is the probability that among the 19 persons who take the test four will pass in the first attempt, five in the second, and the rest of these 19, namely 10, in the third attempt?
2. The records of a cardiology department of a hospital show that patients after their bypass surgery remain in the hospital for five, seven, or 10 days with probability 0.45, 0.35, and 0.20, respectively. What is the probability that of the next 25 patients who go through bypass surgery, 10 will stay for five days, eight for seven days, and the rest of these patients, namely seven, will stay for 10 days?
3. The quality control department of a company has determined that shipments from a certain supplier find four, six, seven, or nine defective items with probabilities 0.55, 0.20, 0.15, and 0.10, respectively. What is the probability that in the next 10 shipments they will find four defective items in each of three shipments, and six defective items in each of the remaining seven shipments?
4. An urn contains six red, eight green, and 11 yellow marbles. A random sample drawn with replacement of 20 marbles is taken. What is the probability that of the 20 marbles seven are red, six green, and seven are yellow.
5. A computer store receives flash drives in shipments consisting of four different memories. A particular shipment contains 100 flash drives of which 30% are of 1 GB, 25% of 2 GB, 25% of 8 GB, and the remaining 20% are of 16 GB memory. A random sample of 20 flash drives is selected from that shipment. What is the probability that the selected sample has six flash drives of 1 GB memory, four of 2 GB, seven of 8 GB, and the remaining have 16 GB memory?
6. A regular die is rolled 12 times. What is the probability of getting two threes, four ones, and three fives?
7. It is believed that the probability of auditing a tax return by the IRS depends on the gross income. Suppose that these probabilities are 0.15, 0.18, 0.27, and 0.40 for the tax returns being audited if their gross incomes are $100K or less, more than $100K but less than $250K, more than $250K but less than $500K, and more than $500K, respectively. What is the probability that among 60 tax returns being audited, 10 have gross income of $100K or less, 15 have gross income of $100K or more but less than $250K, 10 have gross income of $250K or more but less than or equal to $500K, and the remaining 25 have the gross income of $500K or more?

4.8 THE POISSON DISTRIBUTION

4.8.1 Definition and Properties of the Poisson Distribution

We now consider an important probability distribution, the *Poisson distribution*, obtained as a limiting form of the probability function $b(x)$ of the binomial distribution as $n \to \infty$ and $p \to 0$ in such a way that np remains constant. Since $p \to 0$, the Poisson distribution is also known as a probability distribution of *rare events*. Thus, for example, the Poisson distribution may be used to find the probability of the number of occurrences of a particular (rare) event that happens over a specified period of time, or over a specified length of measurement (e.g., the length of an electric wire), or over a specified area or specified volume, when the probability of the event of interest happening is very small. It has application in many areas, in particular as a description of count phenomena.

For example, we may be interested in finding the probability of a number of accidents occurring in a manufacturing plant, the number of patients admitted in a hospital, the number of cars passing through a toll booth, the number of customers entering a bank, or the number of telephone calls received by a receptionist over a specified period of time. Similarly, we may be interested in finding the probability of an electric wire of specified length having a certain kind of defect, the number of scratches over a specified area of a smooth surface, number of holes in a roll of paper, or the number of radioactive particles in a specified volume of space. All these examples have one thing in common: the random variable X denoting the number of occurrences of events that may occur over a specified period of time, length, area, or volume must satisfy the conditions of a process known as the *Poisson process* for us to be able to use the Poisson distribution.

4.8.2 Poisson Process

Let $X(t)$ denote the number of times a particular event occurs randomly in a time period t. Then, these events are said to form a Poisson process having rate λ ($\lambda > 0$) (i.e., for $t = 1, X(1) = \lambda$), if

1. $X(0) = 0$.
2. The number of events that occur in any two nonoverlapping intervals are independent.
3. The average number of events occurring in any interval is proportional to the size of the interval and does not depend on when or where they occur.
4. The probability of precisely one occurrence in a very small interval $(t, t + \delta t)$ of time is equal to $\lambda(\delta t)$, and the probability of two or more occurrences in such a small interval is zero.

4.8.3 Poisson Distribution as a Limiting Form of the Binomial

Consider the binomial probability function,

$$b(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x \frac{(1-p)^n}{(1-p)^x}$$

If $n \to \infty$ and $p \to 0$ in such a way that np remains fixed at a value λ, we obtain ($p = \lambda/n$)

$$\begin{aligned}\lim_{n\to\infty} b(x) &= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \frac{\lambda^x}{n^x} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^x} \\ &= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^x}\end{aligned}$$

$$= \lim_{n\to\infty} 1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\frac{\lambda^x}{x!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^x}$$
$$= \frac{\lambda^x}{x!}e^{-\lambda}$$

The sample space of X is clearly $0, 1, 2, \ldots$, and, of course, e is the base of natural logarithm with value 2.71828. Denoting this limit by $p(x)$, we may write

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0, 1, 2, \ldots \qquad (4.8.1)$$

which is the probability function of the *Poisson distribution.* As the result above suggests, we may use it to approximate $b(x)$ for *large n and small p* as follows:

$$b(x) \approx \frac{e^{-np}(np)^x}{x!} \qquad (4.8.2)$$

The approximation above works well for n large, p small such that $np < 10$. Note that

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\cdots\right) = e^{-\lambda}e^{\lambda} = 1$$

that is, the sum of $p(x)$ over all points in the infinite sample space of X is 1.

Example 4.8.1 (Approximating binomial probability using the Poisson distribution) *Two percent of the screws made by a machine are defective, the defectives occurring at random during production. If the screws are packaged 100 per box, what is the probability that a given box will contain* x *defectives?*

Solution: We assume that the number of screws is produced very large, so we may use the binomial distribution. The probability that the box contains x defectives as given by the binomial distribution is

$$p(x) \approx b(x) = \binom{100}{x}(0.02)^x(1-0.02)^{100-x}; \quad x = 0, 1, \ldots, 100$$

Since $n = 100, p = 0.02$, and $np = 2$, the Poisson approximation to the $b(x)$ above is given by

$$b(x) \approx \frac{e^{-2}(2)^x}{x!}$$

A comparison of $b(x)$ and its Poisson approximation is given in Table 4.8.1.

Table 4.8.1 Values of $b(x)$ and $p(x)$ for $n = 100, p = 0.02, \lambda = 2$.

x	$b(x)$	$p(x)$
0	0.1326	0.1353
1	0.2707	0.2707
2	0.2734	0.2707
3	0.1823	0.1804
4	0.0902	0.0902
5	0.0353	0.0361
6	0.0114	0.0120
7	0.0031	0.0034
8	0.0007	0.0009
9	0.0002	0.0002
10	0.0000	0.0000

Note that if a certain type of wire, for example, is insulated by an enameling process, and if the occurrence of an insulation break follows a Poisson process, then the probability that x insulation breaks will occur in a length L of wire, say $p_x(L)$, is given by

$$p_x(L) = \frac{e^{-\lambda L}(\lambda L)^x}{x!} \tag{4.8.3}$$

where λ is the mean number of insulation breaks per unit length.

Example 4.8.2 (Poisson experiment) *It is known that in a certain enameling process, the number of insulation breaks per yard is 0.07. What is the probability of finding X such breaks in a piece of wire 16 yards long?*

Solution: For a piece of wire 16 yards long, the expected number of insulation breaks is $\lambda L = (0.07) \times 16 = 1.12$. Hence, the probability function for x breaks in a piece of wire 16 yards long is Poisson with mean 1.12, that is

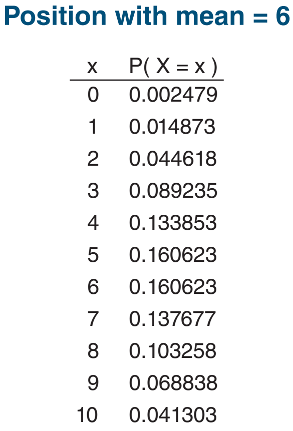
$$p_x(L) = \frac{e^{-1.12}(1.12)^x}{x!}$$

Example 4.8.3 (Using MINITAB and R) *A manufacturing company of car parts found that one of its machines randomly produces some defective parts. Further, the company determined that* X, *the number of defective parts it produces in each shift, is distributed by the Poisson distribution with mean* $\lambda = 3$. *Find using MINITAB and R the probability that it will produce x number of defective parts in the next two shifts, where* $x = 0, 1, 2, \ldots, 10$.

Solution: From equation (4.8.3), it follows that X the number of defective parts the machine will produce in *two shifts* is distributed as Poisson with mean $3 \times 2 = 6$. To find the probabilities $p(X = x), x = 0, 1, 2, \ldots, 10$, proceed as follows:

MINITAB

1. Enter the values 0, 1, 2, . . . , 10 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distributions** > **Poisson**.
3. In the dialog box that appears on the screen, click the circle next to **Probability**.
4. Enter 6 (the value of the mean λL) in the box next to **Mean**.
5. Click the circle next to **Input column** and type C1 in the box next to it.
6. Click **OK**. The desired probabilities will show up in the Session window as:



Position with mean = 6

x	P(X = x)
0	0.002479
1	0.014873
2	0.044618
3	0.089235
4	0.133853
5	0.160623
6	0.160623
7	0.137677
8	0.103258
9	0.068838
10	0.041303

If we want to store these probabilities in a column, say in column C2, then type C2 in the box next to **Optional storage**.

USING R

R has the built in ‘dpois(x, lambda)’ function in ‘stats’ library that can be used to calculate the Poisson probabilities, where x is a nonnegative integer (a quantile), and here, λ is the mean of the Poisson distribution. Probabilities that $x = 0, 1, 2, 3, \ldots, 10$ can be obtained by running the following R code in the R Console window.

```
prob = dpois(c(0:10), lambda=6)

round(prob, 6)
```

We now turn to the moment-generating function (see Section 4.2) of the Poisson distribution random variable X which is given by

$$M_X(t) = exp\{\lambda(e^t - 1)\} \tag{4.8.4}$$

For the derivation of the moment-generating function and some moments, see Section 4.10. The *mean* and *variance* of a Poisson random variable X are given by

$$\boxed{\mu = \lambda, \quad \sigma^2 = \lambda \qquad (4.8.5)}$$

Note that the mean and variance of a Poisson random variable are both equal to λ. For a proof of the results (4.8.5), see Section 4.10.

PRACTICE PROBLEMS FOR SECTION 4.8

1. A random variable X is distributed as Poisson distribution with $\lambda = 3.5$. Use this Poisson distribution (Table A.3) to determine the following probabilities: (a) $P(X < 5)$, (b) $P(2 \leq X \leq 6)$, (c) $P(X > 7)$, (d) $P(X \geq 5)$.
2. A machine in a manufacturing plant has on the average two breakdowns per month. Find the probability that during the next three months it has (a) at least five breakdowns, (b) at most eight breakdowns, (c) more than five breakdowns.
3. The number of defective parts produced per shift can be modeled using a random variable that has the Poisson distribution. Assume that, on average, three defective parts per shift are produced.
   (a) What is the probability that exactly four defective parts are produced in a given shift?
   (b) What is the probability that more than seven defective parts are produced in the next two shifts?
   (c) What is the probability that at the most eight defective parts are produced in the next three shifts?
4. The probability that a woman will die from breast or cervical cancer is 0.00027. Find the probability that of the 10,000 women who are monitored for breast or cervical cancer: (a) three will die from breast or cervical cancer, (b) at most four will die from breast or cervical cancer, or (c) at least two will die from breast or cervical cancer.
5. Based on past records, a retail store manager knows that on average, 30 customers per hour come to the store. Find the probability that in a given five-minute period:
   (a) At least two will come to the store
   (b) Exactly four will come to the store
   (c) At most six will come to the store
6. The number of complaints that a discount store receives on their product per week is a random variable having Poisson distribution with $\lambda = 5$. Find the probability that during any given period of two weeks it will receive: (a) at least 10, (b) exactly nine, (c) at most 12.
7. Suppose that the probability that an insurance company pays out a claim in a given six-month period against a car theft is 0.0003. Find the probability that of the 15,000 insurers against car theft, it will pay out at least 10 claims during any given year.
8. A random variable X is distributed by the binomial distribution with $n = 15, p = 0.1$. Find the following probabilities, first using the binomial distribution and then using the Poisson approximation to the binomial distribution. Compare your results.

(a) $P(X < 6)$
(b) $P(4 \leq X \leq 9)$
(c) $P(X \geq 5)$
(d) $P(X \leq 8)$

4.9 THE NEGATIVE BINOMIAL DISTRIBUTION

The *negative binomial distribution* is applicable whenever an experiment possesses the following characteristics:

1. The experiment consists of a sequence of independent trials.
2. Each trial results in either one of two possible outcomes, usually referred to as success and failure.
3. The probability of success for each trial, p, is constant across the experiment, and consequently the probability, q, of failure is constant throughout the experiment.
4. The experiment continues until a fixed number of successes has been achieved.

Let A denote the event that a trial is a success and $\bar{A}$ the event that a trial is a failure, and suppose that $P(A) = p$ and $P(\bar{A}) = q$. We now determine the probability that the number of trials required to obtain exactly k successes is x. Note that in the binomial distribution, the number of trials is fixed but that the number of successes found in a fixed number of trials is a random variable. In the negative binomial scenario, this is reversed; that is, the number of trials X is a random variable, while the number of successes is fixed.

To determine the probability of k successes at the xth trial, suppose that we let E be the event of obtaining $k-1$ A's in the first $x-1$ trials and F be the event of getting an A on the xth trial. Since trials are independent and since we are interested in the event $E \cap F$, where E and F are independent events, we have

$$P(X = x) = P(E \cap F) = P(E)P(F) \tag{4.9.1}$$

But the probability $P(E)$ in equation (4.9.1) is the binomial probability of $(k-1)$ successes in $(x-1)$ trials, which is, since $(x-1)-(k-1) = x-k$, given by,

$$P(E) = \binom{x-1}{k-1} p^{k-1} q^{x-k} \tag{4.9.2}$$

Also, $P(F) = p$, the probability of success in the xth trial. Thus $p(x)$, the probability function of the negative binomial random variable X, is given by

$$p(x) = P(E)P(F) = \binom{x-1}{k-1} p^k q^{x-k}; \quad x = k, k+1, \ldots \tag{4.9.3}$$

This is called the *negative binomial distribution* because $p(x)$ is the $(x-k+1)$th term obtained in the expansion of $p^k(1-q)^{-k}$, when $(1-q)^{-k}$ (a binomial $(1-q)$ with negative exponent $-k$) is expanded into a series in powers of q, where of course $q = 1-p$. It is interesting also to note that (4.9.3) is sometimes referred to as the *binomial waiting-time distribution*, since the probability function $p(x)$ is simply the probability that one must wait through x independent trials in order to obtain k successes, that is, k A's.

Example 4.9.1 (Negative binomial probabilities) *Two dice are thrown and the sum of the dots obtained on the uppermost faces are recorded. What is the probability that a 7 occurs for the third time on the third throw? On the fourth throw? On the* x*th throw* $(x \geq 3)$*?*

Solution: Consulting Example 3.7.1, we have that $p = p(7) = 6/36 = 1/6$ and that the number of throws x, needed to get $k = 3$ sevens (event A is the event "a seven") has probabilities given by the negative binomial distribution (see (4.9.3)), so that

$$p(x) = \binom{x-1}{2} p^3 q^{x-3} = \binom{x-1}{2} \frac{5^{x-3}}{6^x}; \quad \text{for } x \geq 3$$

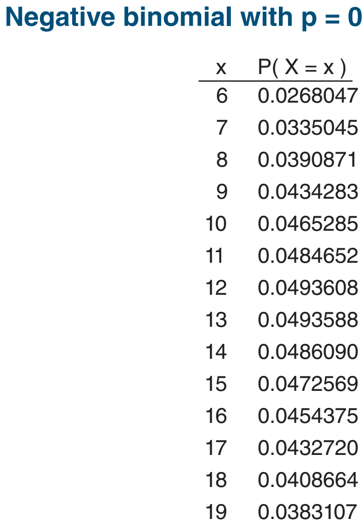
Hence,

$$p(3) = 1/216 = 0.00463, \quad p(4) = 15/1296 = 0.01157, \quad p(5) = 150/7776 = 0.01929$$

We determine the probabilities for $x = 6, 7, \ldots,$ by using MINITAB and R as follows.

MINITAB

1. Enter the values 6, 7, . . . , in column C1.
2. From the Menu bar select **Calc** > **Probability Distributions** > **Negative Binomial**.
3. In the dialog box that appears on the screen, click the circle next to **Probability**.
4. Enter 3 (the number of successes) in the box next to **Number of Events needed** and $0.1667 = 1/6$ (the probability of success) in the box next to **Event Probability**.
5. Click the circle next to **Input column** and type C1 in the box next to it.
6. Click **OK**. The desired probabilities will show up in the Session window as:



```
Negative binomial with p = 0.1667 and r = 3

 x   P( X = x )
 6    0.0268047
 7    0.0335045
 8    0.0390871
 9    0.0434283
10    0.0465285
11    0.0484652
12    0.0493608
13    0.0493588
14    0.0486090
15    0.0472569
16    0.0454375
17    0.0432720
18    0.0408664
19    0.0383107
```

NOTE: X = total number of trials, and in MINITAB k is denoted by r.

If we want to store these probabilities in a column, say in column C2, then type C2 in the box next to **Optional storage**.

USING R

R has the built in ‘dnbinom(x, size, prob)’ function in ‘stats’ library that can be used to calculate the Negative Binomial probabilities, where ‘x’ is the number failures prior to r many successes (‘size’). Also, ‘prob’ is the probability of success in a single trial. To get three successes at the sixth trial, we should have three failures. Therefore, to get the required probabilities as shown in the previously mentioned MINITAB output, we have to input the number of failures $x = 3, 4, 5, 6, \ldots, 16$. Required probabilities can be obtained by running the following R code in the R Console window.

```
probabilities = dnbinom(c(3:16), 3, 0.1667)
round(probabilities, 6)

#R output

[1] 0.026805 0.033505 0.039087 0.043428 0.046528 0.048465 0.049361
[8] 0.049359 0.048609 0.047257 0.045437 0.043272 0.040866 0.038311
```

The *mean* and *variance* of the negative binomial distribution with probability function given in Equation (4.9.3) are

$$\mu = k/p, \quad \sigma^2 = kq/p^2 \tag{4.9.4}$$

As an example, if the aforementioned experiment was repeated a large number of times, the mean wait (k/p) needed to get 3 sevens would be $3/(1/6) = 18$ throws. It can easily be shown that the moment-generating function of the negative binomial distribution is given by

$$M_X(t) = \left(\frac{pe^t}{1 - qe^t}\right)^k \tag{4.9.5}$$

PRACTICE PROBLEMS FOR SECTION 4.9

1. A manufacturing company of wind turbines found that the probability that a turbine is nonconforming is 0.03. Assume that the turbines are conforming or nonconforming independently. Find the probability that the third nonconforming turbine is the 100th turbine manufactured by that company.

2. Consider a sequence of independent Bernoulli trials with probability of success being 0.25. Determine the probability that (a) the fifth success occurs at the 16th trial, (b) the fourth success occurs at the 10th trial.
3. Suppose that independent trials are carried out, each resulting in a success with probability 0.6. What is the probability that the ninth success occurs in the 20th trial?
4. Referring to Problem 3, find the probability that the 14th success occurs before the 12th failure.
5. The night shift in a manufacturing plant is known to produce 10% of its items defective. A quality control engineer inspects all the items that were manufactured in a given night shift. What is the probability that to find the fifth defective item, the engineer will have to inspect at least 31 items?
6. The probability that a baseball player hits a home run in any one inning is 0.30. What is the probability that he will hit a second home run in the eighth inning if he bats every inning?
7. A quality control engineer in a manufacturing company detects that a recently purchased machine produces 5% of parts that do not meet the specifications. What is the probability that the third part that does not meet the specifications is the 40th part produced by that machine?

4.10 SOME DERIVATIONS AND PROOFS (OPTIONAL)

This section is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

4.11 A CASE STUDY

Case Study (*Reducing errors by vendors processing credit-card applications*)[1] This case study pertains to the problem of reducing errors by vendors who process credit-card applications for a large credit-card bank. Each vendor has workers who physically process the applications by checking them for completeness and then entering the information into a computer. In this case study, the quality control sampling plan proceeds by selecting a sample of size 50 completed credit-card applications every day from each vendor. A processed credit-card application is labeled defective (nonconforming) if there is any error on the application. In this case study, the authors obtained the data for four particular vendors for period of time ranging from 107 to 207 days. The data for this case study are available on the book website: www.wiley.com/college/gupta/statistics2e.

(a) Find the average number of nonconforming applications per sample for each of the four vendors.

(b) Using the results you obtained in part (a) and employing the Poisson distribution, find the probabilities for each vendor of finding $X = 0, 1, \ldots, 5$ nonconforming applications.

[1] Source: Lucas, Davis, and Saniga (2006) and Saniga, Davis, and Lucas (2009). Used with permission.

(c) Compare the probabilities you determined in part (b) and comment on the quality of work of these vendors.

4.12 USING JMP

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Suppose that a lot of 50 fuses, of which seven are known to be defective, is available for sampling. It is proposed to draw 10 fuses at random (without replacement) and test them. What is the probability that such random samples of 10 will contain 0, 1, $2, \ldots$, 7 defective fuses?
2. A lot contains 30 items, six of which are defective. What is the probability that a random sample of five items from the lot will contain no defective items? No more than one defective? More than two defectives? (Assume that sampling is without replacement.)
3. In rolling five true dice, find the probability of obtaining at least one ace, exactly one ace, exactly two aces. (Here, ace implies one point.)
4. If the probability of hitting a target is 0.2 and 10 shots are fired independently, what is the probability that the target will be hit at least once? At least twice?
5. What is the probability of drawing a 13-card hand containing no aces, kings, queens, or jacks?
6. Suppose that 5% of the aspirins pressed by a certain type of machine are chipped. The tablets are boxed 12 per box. What percent of the boxes would you estimate:
   (a) To be free of chipped tablets?
   (b) To have not more than one chipped tablet?
   (c) To have exactly x chipped tablets?
7. Suppose that 13 cards are dealt from a thoroughly shuffled deck of ordinary playing cards.
   (a) What is the probability of getting x spades?
   (b) What is the probability of getting y hearts? Describe the sample space of y.
   (c) What is the probability of getting x spades and y hearts? Describe the sample space of $(x,\ y)$.
8. What is the probability of throwing two heads three times in four throws of five coins?
9. It is known that 0.0005% of the insured males die from a certain kind of accident each year. What is the probability that an insurance company must pay off on more than three of 10,000 insured against such accidents in a given year?
10. A bag of grass seed is known to contain 1% weed seeds. A sample of 100 seeds is drawn. Find the probabilities of 0, 1, 2, $3, \ldots$, 7 weed seeds being in the sample.

11. A process for making plate glass produces an average of four "seeds" (small bubbles) scattered at random in the glass per 100 ft^2. With the use of the Poisson distribution, what is the probability that
    (a) A piece of plate glass five ft by 10 ft will contain more than two seeds?
    (b) Six pieces of plate glass five ft by five ft will all be free of seeds?

12. Small brass pins are made by company ABC. Of the pins manufactured, 2% are undersized, 6% are oversized, and 92% are satisfactory. The pins are boxed 100 per box. A box is taken at random. Write down the expression for the probabilities of the following events:
    (a) The box contains x satisfactory pins and y undersized pins, the remaining pins being oversized.
    (b) The box contains no undersized pins.
    (c) The box contains only satisfactory pins.

13. Suppose that 10 people each throw two coins. What is the probability that:
    (a) Three people throw two heads, three people throw two tails, and four people throw one head and one tail?
    (b) No one threw a head and a tail?

14. Two coins are tossed n times. Find the probability of x, the number of times no heads appear; y, the number of times one head appears; and z, the number of times two heads appear $(x + y + z = n)$.

15. An urn contains 10 white and 20 black balls. Balls are drawn one by one, without replacement, until five white ones have appeared. Let X be the number of draws necessary to find five white balls. What is the sample space of X? Find an expression for the probability of the event $X = x$.

16. Suppose that a lot of 10,000 articles has 200 defectives and that a random sample of 100 articles is drawn from the lot (without replacement).
    (a) What is the probability of getting exactly x defectives in the sample?
    (b) Determine the binomial approximation for the probability in (a).
    (c) Determine the Poisson approximation for the probability in (a).

17. The fraction of articles turned out by a machine that is defective is equal to p. The defectives occur "at random" during production. The articles are boxed m per box and cartoned n boxes per carton (assume production numbers are large).
    (a) If a box of articles is taken at random, what is the probability it contains exactly x defectives?
    (b) If a carton is taken at random, what is the probability that exactly y of its boxes will be free of defectives?

18. In Problem 17, what is the probability that
    (a) The machine will produce k good (nondefective) articles before it turns out a defective?
    (b) The machine has to turn out a total of x articles in order to produce exactly k good ones?

19. If the probability is 0.6 that an engineer will pass the six-sigma black belt test in the first attempt, use the formula for the binomial distribution to find the probability that 6 of 10 engineers taking that test will pass in the first attempt.

20. A lot of N articles has d defectives. If articles are taken at random from the lot one at a time, what is the probability, assuming sampling without replacement, that
    (a) Exactly x articles have to be examined to find the first defective?
    (b) Exactly x articles have to be examined to find the dth (last) defective?

21. In testing a relay, suppose that the probability is p that it fails to make satisfactory contact in a single trial and that p remains unchanged over a very large number of trials. By assuming the outcomes of successive trials to be independent, what is the probability that
    (a) x trials have to be made to obtain the first failure?
    (b) x trials have to be made to obtain the kth failure?

22. In 1970s, the proportion of successful launches at a test missile site has been 0.85. Suppose that an experiment is planned that requires three successful launches. What is the probability that exactly x attempts will be necessary? exactly five? exactly seven? fewer than six?

23. In Problem 22, suppose that another experiment requires four successful launches. What is the probability that six attempts will be required? What is the probability that x attempts will be required?

24. A device fails to operate on a single trial with probability p. Let X be a random variable denoting the number of trials required to obtain a total of k failures. If results of successive trials are independent, show that the probability function of X is given by

$$p(x) = \binom{x-1}{k-1} p^k q^{x-k}; \quad x = k, k+1, \ldots$$

Show that the mean and variance of X are given by

$$E(X) = \frac{k}{p}, \quad Var(X) = \frac{kq}{p^2}$$

25. In the Problem 24 above, suppose $k = 1$. Show that if Y is a random variable denoting the number of trials required to obtain one failure, then the p.f. of Y is

$$p(y) = pq^{y-1}, \quad y = 1, 2, \ldots$$

and that its mean and variance are $1/p$ and q/p^2, respectively. This distribution is called the *geometric distribution*. Show that $\sum_{y=1}^{\infty} p(y) = 1$.

26. Referring to Problem 25, show that $P(Y > s + t | Y > s) = P(Y > t) = \sum_{y=t+1}^{\infty} pq^{y-1}$. (This result implies that the geometric distribution has no memory, for if the event of a failure has *not* occurred during the first s trials, then the probability that a failure will not occur in the next t trials is the same as the probability that it will not occur in the first t trials. In other words, the information that a failure has not occurred in the first s trials is “forgotten” in the subsequent calculations.)

27. A lot contains N articles of which Np are defective. Articles are drawn successively at random and without replacement until k defectives are drawn. Let X be a random variable denoting the number of articles that must be drawn to achieve this objective.

Show that the p.f. of x is given by

$$p(x) = \frac{\binom{x-1}{k-1}\binom{N-x}{Np-k}}{\binom{N}{Np}}; \quad x = k, k+1, \ldots$$

and that $E(X) = \dfrac{k(N+1)}{Np+1}$.

28. By using the moment-generating function (4.6.4) of a random variable X having the binomial distribution (4.6.2), show that the mean and variance of X are np and npq, respectively.

29. In testing electric bulbs for a certain kind of projector, it is found that 40% of the bulbs burn out before the warranty period. Suppose that the engineering department of a school buys 12 such bulbs. Find the probability that:
   (a) Between four to six bulbs (inclusive) burn out before the warranty period.
   (b) More than five bulbs burn out before the warranty period.
   (c) Fewer than eight bulbs burn out before the warranty period.
   (d) No bulb burns out before the warranty period.

30. Let X be a random variable distributed as binomial distribution with $n = 25$ and $p = 0.35$. Find the mean, variance, and the standard deviation of the random variable X.

31. The drug Xanax is used to control an anxiety problem. However, it is believed that 70% of the users get addicted to the drug. Suppose that we take a random sample of 15 Xanax users and find the number of persons addicted to Xanax. Find the probability that:
   (a) More than 10 are addicted to Xanax.
   (b) Fewer than eight are addicted to Xanax.
   (c) Between 10 to 12 inclusive are addicted to Xanax.

32. A box of 100 computer chips contains eight defective chips. Suppose that a random sample of size 10 chips is selected without replacement from that box. Find the probability that the sample had
   (a) At least one defective chip.
   (b) All defective chips.
   (c) Nine defective chips.
   (d) No defective chips.

33. In Problem 32, let X denote the number of defective chips. Find the mean, variance, and the standard deviation of the random variable X.

34. An engineering club consists of five seniors and seven juniors. Suppose that five club members are selected randomly to form a committee. Find the probability that
   (a) The committee has at least two juniors.
   (b) The committee has three or more seniors.
   (c) The committee has no more than two seniors.
   (d) The committee has no junior.

35. An insurance company discovered that three policyholders out of every 1000 insured against a particular kind of accident file a claim every year. Suppose that the company

has 2000 persons who are insured against that kind of accident. Find the probability that

(a) During a given year at least four will file the claim.
(b) No more than 10 will file the claim.
(c) Between five to eight (inclusive) will file the claim.
(d) Fewer than two will file the claim.
(e) More than two will file the claim.

36. A programmer makes two wrong entries every hour, on the average. Find the probability that during the next five hours she will make
(a) Fewer than eight wrong entries.
(b) At least four wrong entries.
(c) Between three to five (inclusive) wrong entries.
(d) More than one wrong entry.

37. On average, the number of customers arriving per every ten minutes at a teller's window in a bank is four. Find the probability that during the next 10 min:
(a) At least five customers will arrive at that teller's window.
(b) No more than two customers will arrive at that teller's window.
(c) Between two to six (inclusive) customers will arrive at that teller's window.
(d) Less than six customers will arrive at that teller's window.

38. Indicate which of the following experiments can be studied using a binomial model. Justify your answer.
(a) Drawing five ball-bearings with replacement from a box containing 25 ball-bearings, 10 of which are of diameter 10 mm and 15 of diameter 20 mm, and observing the diameters of the drawn ball-bearings.
(b) Selecting randomly four engineers to be on a contract negotiating team from a group of 50 engineers, 20 of whom are six sigma green belts, and observing how many of selected engineers are six sigma green belts.
(c) A fair die is rolled and the number that turns up is observed.
(d) Selecting a manufacturing company from the midwestern part of the United States and observing whether its annual revenues are more than \$1 billion or not when it is known that 30% of all manufacturing companies in that region have annual revenues totaling more than \$1 billion.

39. Just before the 2006 midterm elections of the United States, one of the polling agencies found that 60% of the voters were against the Iraq war. Assume that this result is valid for all the voters in the entire country. Using the binomial distribution table (Table A.2), compute the probability that in a random sample of 20 American voters, the number of those against the Iraq war are
(a) At least five.
(b) At the most seven.
(c) More than five but less than 10.
(d) Exactly eight.
(e) Less than or equal to nine.

40. A six-sigma green belt quality control engineer found that on average, batches of 500 computer chips have exactly two defective chips.
(a) Using the formula for the Poisson distribution, determine the probability that a box of 1000 chips will have exactly 5 defective chips.

(b) Using the Poisson distribution (Table A.3), compute the probability that a box of 1000 chips will have (i) more than five defective chips, (ii) at the most six defective chips, (iii) between four and eight (inclusive) defective chips.

41. A batch of 500 car batteries is scheduled to be shipped if a random sample of 20 from the batch has two or fewer defective batteries. If it is known that there are 60 defective batteries in the batch, find the probability that the batch will be shipped.

42. The number of patients admitted in an emergency room of a metropolitan hospital can be modeled as a Poisson random variable. Assume that on the average, five patients are admitted every hour.
    (a) What is the probability that exactly four patients are admitted in the next one hour?
    (b) What is the probability that more than seven patients are admitted in the next two hours?
    (c) What is the probability that at the most eight patients are admitted in the next 90 min?
    (d) What is the probability that more than five, but less than 10, patients are admitted in the next two hours?

43. Which of the following functions are valid probability functions? Explain.
    (a) $p(x) = x/20; x = 1, 2, 3, 4, 5, 6$ and zero elsewhere.
    (b) $p(x) = x^2/140; x = 1, 2, 3, 4, 5, 6, 7$ and zero elsewhere.
    (c) $p(x) = (x - 3)/5; x = 2, 3, 4, 5, 6$ and zero elsewhere.

44. For the functions that are valid probability functions in Problem 43, find the mean and the variance of X.

45. Determine the value of the constant c such that the following functions are valid probability functions:
    (a) $p(x) = cx/20; x = 1, 2, 3, 4, 5, 6$ and zero elsewhere.
    (b) $p(x) = c(x^2 + 1); x = 1, 2, 3, 4, 5$ and zero elsewhere.
    (c) $p(x) = c(x - 1); x = 1, 2, 3, 4, 5, 6$ and zero elsewhere.

46. Refer to Problem 45. In each case, determine the mean and the variance of X.

47. Determine the mean and the variance of the following probability functions:
    (a) $p(x) = x/21; x = 1, 2, 3, 4, 5, 6$ and zero elsewhere.
    (b) $p(x) = (x^2 - 1)/50; x = 1, 2, 3, 4, 5$ and zero elsewhere.

48. Let X be a random variable having the uniform distribution on $x = 1, 2, \ldots, N$. Find the mean and the variance of X.

49. Let X be a random variable that is Bernoulli distributed with parameter p. Find the moment-generating function of X. Then, use the moment-generating function to find the mean and the variance of X.

50. Let the random variable X have a discrete uniform distribution on the integers $0 \leq x \leq 50$. Determine the mean and the variance of X.

51. Refer to Problem 48. Find the mean and the variance of $X + c$ (c is constant) and comment.

52. In analyzing a large data set, a life insurance company estimated the probability that a person in the 70-80 years of age group dies due to natural causes in any given year

as 0.000002. If the company has 500,000 insurers in that age group, determine the probability that due to death from natural causes, the company will have to pay off during any given year:

(a) At least three claims.
(b) No more than four claims.
(c) Between three and five (inclusive) claims.

53. A circuit board of a very complex electronic system has 500 soldered joints. The probability that a joint becomes loose in one year use of the circuit board is 0.001.
    (a) What is the probability that in one year three joints become loose?
    (b) What is the probability that in one year at least two joints become loose?
    (c) The circuit board becomes dysfunctional if any joint becomes loose. Find the probability that in two years the circuit board becomes dysfunctional.

54. The probability that a part manufactured at a plant being defective is 0.001. On a given day, the plant manufactured 10,000 parts. Find the following probabilities:
    (a) At least five parts are defective.
    (b) No more than eight parts are defective.
    (c) Between four and nine (inclusive) parts are defective.

55. A quality control engineer is interested to find how many parts he/she needs to inspect to detect the first defective part. If the probability that a randomly selected part is defective is p and X is the number of parts inspected needed to detect the first defective part, then X is a random variable that is distributed by the geometric distribution.
    (a) Determine the probability distribution of the random variable X.
    (b) Determine the expected value and the variance of X.
    (c) Determine $P(X = 20)$, given that $p = 0.07$.

56. The random variable X is distributed by the geometric distribution with $p = 0.05$. Determine the following probabilities: (a) $P(X \geq 24)$, (b) $P(15 \leq X \leq 30)$, (c) $P(X > 28)$.

57. A pool of 15 applicants for the position of a manager for a firm consists of 10 applicants holding master's degrees and five holding PhD degrees. An interviewer randomly selects eight applicants to interview. Determine the probabilities of the following events:
    (a) He/she selects three applicants with a PhD degree.
    (b) He/she selects at least two applicants with a PhD degree.
    (c) He/she selects no more than two applicants with a PhD degree.

58. Of all customers buying cars at a car fair, 60% buy an American car. Let a random variable X represent the number of customers who bought an American car out of a total 50 cars sold at the fair.
    (a) Describe the probability distribution of the random variable X.
    (b) Determine the expected value and the variance of X.
    (c) Determine the probability $P(X \leq 30)$.
    (d) Determine the probability $P(15 \leq X \leq 25)$.
    (e) Determine the probability $P(X > 20)$.

Chapter 5

CONTINUOUS RANDOM VARIABLES AND SOME IMPORTANT CONTINUOUS PROBABILITY DISTRIBUTIONS

The focus of this chapter is a discussion of some important continuous probability distributions.

Topics Covered

- Continuous random variables and their probability distributions
- Determination of cumulative distribution functions (c.d.f.'s) from probability density functions (p.d.f.'s)
- Determination of cumulative probabilities for different probability distributions
- Determination of the mean and variance of different continuous probability distributions, including the normal, exponential, gamma, and Weibull distributions
- Determination of the cumulative probabilities for different probability distributions using the statistical packages MINITAB, R, and JMP
- Approximation of the binomial and Poisson distributions by the normal distribution
- Determination of the mean and the variance of linear functions of independent normal random variables
- Test of normality
- Some reliability theory related probability models: lognormal, exponential, gamma, and Weibull distribution

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the difference between discrete and continuous random variables.
- Understand various important continuous distributions and apply them to determine probabilities in real-world problems.
- Determine approximate probabilities of discrete random variables using the normal distribution.
- Determine the mean and the variance of continuous random variables using the usual techniques, and/or moment-generating functions (m.g.f.s).
- Understand various probability models extensively used in reliability theory.
- Apply statistical packages MINITAB, R, and JMP to calculate probabilities when using different probability models of continuous random variables.

5.1 CONTINUOUS RANDOM VARIABLES

In Chapter 4, we discussed discrete random variables, for which the sample space contains a countable number of values. In many applications, the sample space contains an infinite number of uncountable values. In such cases, we cannot associate probabilities with the different possible values that the random variable can take on in the same way as for a discrete random variable. For this scenario, we develop another class of random variables called *continuous random variables.* In dealing with continuous random variables, we are always interested in calculating the probability of random variables taking any value in an interval rather than taking on any individual values. For example, consider a problem of finding the probability that a technician takes certain time to finish a job. Consider a random variable X denoting the time (in hours) taken by the technician to finish the job. Then, we may be interested in finding, for example, the probability $P(3.0 \leq X \leq 3.5)$, that is, the probability that she takes between three and three-and-one-half hours to finish the job, rather than finding the probability that she will take *exactly* three hours, 10 minutes, and 15 seconds to finish the job. In this case, the event associated with completing the job in *exactly* three hours, 10 minutes, and 15 seconds is virtually impossible. In other words, as we will see, the probability of such an event is zero.

Definition 5.1.1 A random variable X is said to be of the *continuous type* if its sample space consists of all values in an interval or in many intervals.

Suppose S is a sample space with an uncountable number of values, and let E_x be the event consisting of all values e for which $X(e) \leq x$ or expressed more briefly the set $\{e : X(e) \leq x\}$. We assign a probability, say $F(x)$ to E_x, for every value of x. That is, we have

$$P[X(e) \leq x] = P(E_x) = F(x).$$

Clearly, $F(x)$, the so-called (c.d.f.) of X, is such that

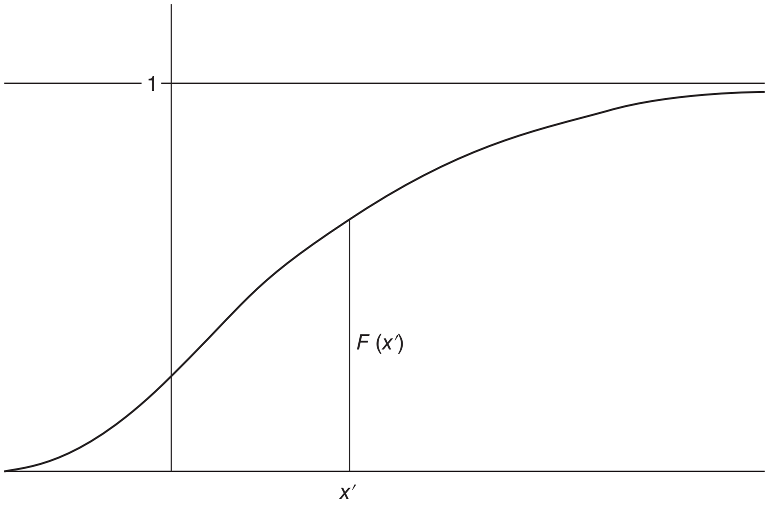


Figure 5.1.1 Graph of the cumulative distribution for a continuous variable.

$$0 \leq F(x) \leq 1 \tag{5.1.1a}$$

$$\text{If } \ x_1 < x_2, \ \text{ then } \ F(x_1) \leq F(x_2) \tag{5.1.1b}$$

$$F(-\infty) = 0, \quad F(+\infty) = 1 \tag{5.1.1c}$$

If $F(x)$ is continuous in x, we say that X is a continuous random variable and that $F(x)$ is the *c.d.f.* of X. We have seen in Section 4.1 that if X is a discrete random variable, then the c.d.f. $F(x)$ is a step function, as shown in Figure 4.1.2.

In the case of a continuous random variable X, the graph $F(x)$ is as illustrated in Figure 5.1.1; that is, it has no vertical jumps. For any point x', $F(x')$ represents the total amount of probability "smeared" along the x-axis to the left of x'.

If $F(x)$ has a derivative $f(x)$, then $f(x)$ is nonnegative and is called the *p.d.f.* of the random variable X. The relationship between $F(x)$ and $f(x)$ is as follows:

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)dx \tag{5.1.2}$$

Note that the p.d.f. of a continuous random variable possesses the following properties:

$$f(x) \geq 0 \tag{5.1.3a}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{5.1.3b}$$

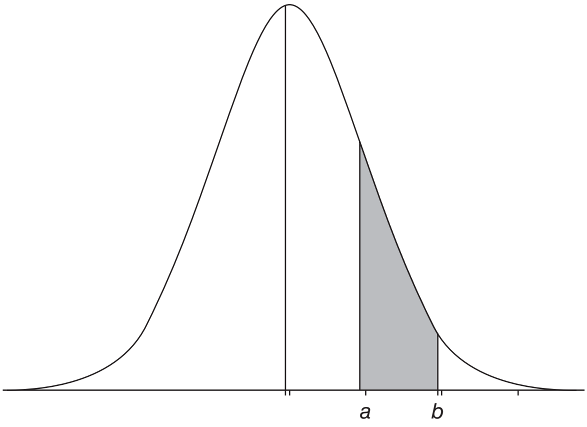


Figure 5.1.2 The probability $P(a \leq X \leq b)$ that the random variable X falls in an interval $[a, b]$ is represented by the shaded area.

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{5.1.3c}$$

$P(a \leq X \leq b) =$ area under the density function $f(x)$ within the interval (a, b)

The mathematical expression in (5.1.3b) represents the total area enclosed by the probability density curve and the x-axis. That is, the total area, which represents the total probability, is equal to one. The probability $P(a \leq X \leq b)$ that the random variable X falls in an interval $[a, b]$ is the shaded area shown in Figure 5.1.2.

If in Figure 5.1.2 we take $a = b$, then the shaded area is zero, which implies that $P(X = a) = P(X = b) = 0$. That is, the probability of the continuous random variable taking any exact value is zero. This fact leads to an important result in that it does not matter if the endpoints of an interval are included or not while calculating the probability that the continuous random variable X falls in an interval (a, b). Hence, if X is a continuous random variable, then

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b) \tag{5.1.4}$$

So far, we have had a very general discussion about the probability distributions of continuous random variables. In this chapter, we discuss some special continuous-probability distributions that we encounter frequently in applied statistics and their properties.

5.2 MEAN AND VARIANCE OF CONTINUOUS RANDOM VARIABLES

5.2.1 Expected Value of Continuous Random Variables and Their Functions

Suppose a continuous random variable X has a p.d.f. $f(x)$. The expectation $E(X)$ of the random variable X is defined as follows:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu \tag{5.2.1}$$

which may be interpreted as the center of gravity or first moment of the probability density function $f(x)$. $E(X)$ is often referred to as the *mean value* of X or simply the *mean* of X. The variance of a continuous random variable X is defined as

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \sigma^2 \tag{5.2.2}$$

As in the discrete case, there is a more convenient expression for the calculation of σ^2, given by

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 \tag{5.2.3}$$

The standard deviation of a random variable X is defined as $SD(X) = +\sqrt{Var(X)} = \sigma$.

Example 5.2.1 *Suppose a continuous random variable* X *has the p.d.f. given by*

$$f(x) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ 2x/R^2 & \text{if} \quad 0 < x \le R \\ 0 & \text{if} \quad x > R \end{cases}$$

Then, using (5.2.1), we have that the mean of X is given by

$$\mu_x = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

so that

$$\mu_x = E(X) = \frac{2}{R^2}\int_0^R x^2 dx = \left(\frac{2}{R^2}\right)\left[\left(\frac{x^3}{3}\right)\Big|_0^R\right] = \frac{2R}{3}$$

Now, using (5.2.3), we find that the variance σ_x^2 of X is given by

$$\sigma_x^2 = \frac{2}{R^2}\int_0^R x^3 dx - \left(\frac{2R}{3}\right)^2 = \left(\frac{2}{R^2}\right)\left[\left(\frac{x^4}{4}\right)\Big|_0^R\right] - \frac{4R^2}{9} = \frac{R^2}{18}$$

Note that the integration in Example 5.2.1 was straightforward. Sometimes integration by parts is necessary, as in the following example.

Example 5.2.2 (Exponential distribution) *Suppose a random variable* X *has the p.d.f. $f(x)$ given by*

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0 \tag{5.2.4}$$

and zero otherwise. Then, the mean of X is

$$\mu_x = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} x\lambda e^{-\lambda x}dx$$

We see that integration by parts is necessary. To this end, suppose we set $u = x$ and $dv = \lambda e^{-\lambda x}dx$. We then have that $du = dx$ and $v = -e^{-\lambda x}$. Using the well-known formula for *integration by parts*, $\int u\, dv = uv - \int v\, du$, the reader may verify that

$$\mu_x = 1/\lambda \tag{5.2.5}$$

Further, using (5.2.3), the variance of X is

$$\sigma_x^2 = Var(X) = \int_0^{\infty} x^2\lambda e^{-\lambda x}dx - (1/\lambda)^2$$

The reader may verify that two applications of integration by parts yield

$$\sigma_x^2 = Var(X) = 2(1/\lambda)^2 - (1/\lambda)^2 = 1/\lambda^2 \tag{5.2.6}$$

That is, the mean and variance of the random variable X, which has a distribution whose p.d.f. is given by (5.2.4), are $1/\lambda$ and $1/\lambda^2$, respectively. The distribution in (5.2.4) is called the *exponential distribution* and will be discussed later in this chapter. Both the mean and the standard deviation of this distribution is equal to $1/\lambda$ and is a unique property of this distribution.

Certain Properties of an Expected Value and Variance of a Random Variable

Using elementary properties of integrals, we have, in general, that (5.2.7) holds:

$$E(cX) = cE(X), \quad Var(cX) = c^2 Var(X) \tag{5.2.7}$$

In Section 4.2, we discussed certain results about expected values of a discrete random variable and its functions. Similar results are also valid for a continuous random variable and its functions. Below we state these results.

Theorem 5.2.1 *Suppose* c *is a constant and* X *is a continuous random variable that is distributed with p.d.f.* $f(x)$. *Then,*

$$E(c) = c \tag{5.2.8}$$

Proof: $E(c) = \int_{-\infty}^{\infty} cf(x)dx = c\int_{-\infty}^{\infty} f(x)dx = c \times 1 = c$ □

Theorem 5.2.2 *Let* c *be a constant and* $g(x)$ *be a function of a continuous random variable* X *that is distributed with p.d.f.* $f(x)$. *Then,*

$$E(cg(X)) = cE(g(X)) \tag{5.2.9}$$

Proof: Using Equation (5.2.7), we have

$$E(cg(X)) = \int_{-\infty}^{\infty} cg(X)f(x)dx = c\int_{-\infty}^{\infty} g(X)f(x)dx = cE(g(X))$$ □

Theorem 5.2.3 *Let* $g_i(x), i = 1, 2, \ldots, n$ *be* n *functions of a continuous random variable* X *that is distributed with p.d.f.* $f(x)$. *Then,*

$$E\left[\sum_{i=1}^{n} g_i(X)\right] = \sum_{i=1}^{n} E[g_i(X)] \tag{5.2.10}$$

Proof:

$$E\left[\sum_{i=1}^{n} g_i(X)\right] = \int_{-\infty}^{\infty}\left[\sum_{i=1}^{n} g_i(X)\right] f(x)dx = \sum_{i=1}^{n}\left[\int_{-\infty}^{\infty} g_i(X)f(x)dx\right] = \sum_{i=1}^{n} E[g_i(X)].$$ □

5.2.2 The Moment-Generating Function and Expected Value of a Special Function of *X*

Referring to (5.2.9) with $c = 1$, if we set $g(X) = e^{Xt}$ and find its expectation, we obtain

$$E(e^{Xt}) = \int_{-\infty}^{\infty} e^{xt} f(x)dx \tag{5.2.11}$$

The function of t so obtained is called the m.g.f. of the random variable X and is denoted by $M_X(t)$, as recorded below in (5.2.12).

Moment-generating function of a random variable X:

$$M_X(t) = E(e^{Xt}) \tag{5.2.12}$$

Note that $M_X(t)$ can be written as

$$\begin{aligned} M_X(t) &= E\left(1 + Xt + \frac{X^2t^2}{2!} + \cdots + \frac{X^kt^k}{k!} + \cdots\right) \\ &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \cdots + \frac{t^k}{k!}E(X^k) + \cdots \end{aligned} \tag{5.2.13}$$

We then have

$$M_X(t) = 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \cdots + \frac{t^k}{k!}\mu_k' + \cdots \tag{5.2.14}$$

and we note that if the moments about the origin $\mu_1', \mu_2', \ldots$ are all finite, then the coefficient of $t^k/k!$ in the expansion of the m.g.f. is the kth moment of X about the origin. If we differentiate $M_X(t)$ k times, we obtain (assuming differentiability under integral signs),

$$\frac{d^k}{dt^k}M_X(t) = E(X^k e^{Xt}) \tag{5.2.15}$$

and if we then set $t = 0$, we find the result (5.2.16) below.

Expression for obtaining moments about the origin using the moment-generating function:

$$\frac{d^k}{dt^k}M_X(0) = E(X^k) = \mu_k'. \tag{5.2.16}$$

We now present some important properties of the m.g.f. of X, $M_X(t)$. First, if we are interested in the random variable cX, where c is any constant, then, by definition,

$$M_{cX}(t) = E(e^{cXt}) = E(e^{Xct}) = M_X(ct) \tag{5.2.17}$$

Second, the m.g.f. for $X + a$, where a is a constant, is

$$M_{X+a}(t) = E(e^{(X+a)t}) = E(e^{at}e^{Xt}) = e^{at}M_X(t) \tag{5.2.18}$$

Note that (5.2.18) enables us to find moments of X about its mean. If we set $a = -\mu$, we have

$$M_{X-\mu}(t) = E(e^{(X-\mu)t}) \tag{5.2.19}$$

A straightforward differentiation gives the result in (5.2.20).

Expression for obtaining central moments using the moment generating function

$$\frac{d^k}{dt^k}M_{X-\mu}(0) = \mu_k. \tag{5.2.20}$$

We now state an important theorem. The proof of this theorem is beyond the scope of this book.

Theorem 5.2.4 *If two random variables* X *and* Y *have the same m.g.f.* $M(t)$*, then their c.d.f.'s are identical. We then say that* X *and* Y *have the same distribution.*

PRACTICE PROBLEMS FOR SECTIONS 5.1 AND 5.2

1. Suppose that the p.d.f. of a random variable X is $f(x) = 2e^{-2x}$ for $x > 0$. Determine the following probabilities:
   (a) $P(X > 4)$, (b) $P(X < 5)$, (c) $P(2 < X < 7)$, (d) $P(X = 5)$.
2. Determine the value of c such that $f(x) = \dfrac{c}{x^2}$ for $x > 1$ represents the p.d.f. of a random variable X.
3. The probability function of the amount of soft drink in a can is $f(x) = 4cx$ for $11.5 < X < 12.5$ oz. Determine the value of c such that $f(x)$ represents a p.d.f.. Then, find the following probabilities:
   (a) $P(X > 11.5)$, (b) $P(X < 12.25)$, and (c) $P(11.75 < X < 12.25)$.
4. In Problem 3, find the mean μ and the variance σ^2 of the random variable X.
5. The lifetime X (in units of 10 years) of a certain component of a home-heating furnace is a random variable with p.d.f.

$$f(x) = \begin{cases} cx^2(1-x), & \text{if } 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

   Determine the value of c. Then find the probability that the life of such a component is more than eight years.
6. Refer to Problem 5. Find the mean μ and the variance σ^2 of the random variable X.

7. Suppose a random variable X has the p.d.f. $f(x)$ given by

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0$$

   Find the m.g.f. of the random variable X and use your result to find the mean μ and the variance σ^2 of the random variable X.
8. The amount of time X (in minutes) by which the departure of a flight is rescheduled has the probability distribution

$$f(x) = \begin{cases} c(100 - x^2), & -10 < x < 10 \\ 0, & \text{otherwise} \end{cases}$$

   Find the value of c. Then, find the following:
   (a) The mean μ and standard deviation σ of X.
   (b) The mean and standard deviation of $60X$. This gives the mean and standard deviation in seconds.
   (c) The mean and standard deviation of $X/60$. This gives the mean and standard deviation in hours.
9. Refer to Problems 3 and 4 above and find the following probabilities:
   (a) $P(\mu - \sigma < X < \mu + \sigma)$
   (b) $P(\mu - 2\sigma < X < \mu + 2\sigma)$
   (c) $P(\mu - 3\sigma < X < \mu + 3\sigma)$

5.3 CHEBYSHEV'S INEQUALITY

Sometimes it is desirable to have a notion about how much probability there is in the tails of a probability function or a p.d.f. A theorem that gives an upper bound for such probabilities may be stated as in (5.3.1) below.

Theorem 5.3.1 *If* X *is a random variable having finite mean* μ *and variance* σ^2*, then*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{5.3.1}$$

where $k > 1$.

The proof of this theorem is not given here but is available on the book website at: www.wiley.com/college/gupta/statistics2e

This inequality (5.3.1) is called Chebyshev's inequality. Note that the theorem states that no matter what distribution a random variable X may have, if its mean μ and variance σ^2 are finite, the total amount of probability lying in the two tails $(-\infty, \mu - k\sigma]$ and $[\mu + k\sigma, +\infty)$ is no greater than $1/k^2$, which implies that the amount of probability lying in the interval $(\mu - k\sigma, \mu + k\sigma)$ is greater than or equal to $1 - 1/k^2$.

Example 5.3.1 (Applying Chebyshev's inequality) *If a sample size* n *is taken from a lot of* N *items containing 10% defectives, show by using the Chebyshev's inequality that the* *probability exceeds 0.99 that the number of defectives in the sample differs from $n/10$ by not more than $3\sqrt{n}\sqrt{(N-n)/(N-1)}$.*

Solution: We know from Section 4.4 that, if D is the number of defectives in the sample and p is the fraction of defectives in the lot, where $p = 1/10$ in this example, then

$$E(D) = \frac{n}{10} \quad \text{and} \quad Var(D) = \frac{9n}{100}\left(\frac{N-n}{N-1}\right)$$

Thus,

$$\mu_d = \frac{n}{10}, \sigma_d = \frac{3}{10}\sqrt{n}\sqrt{(N-n)/(N-1)}$$

and, since $1 - 1/k^2 = 0.99$, or $k^2 = 100$, we have that

$$k\sigma_d = 3\sqrt{n}\sqrt{(N-n)/(N-1)},$$

Since $k = 10$. Hence, it follows from (5.3.1) that

$$P\left(\left|D - \frac{n}{10}\right| < 3\sqrt{n}\sqrt{\frac{N-n}{N-1}}\right) \geq 1 - \frac{1}{100} = 0.99$$

thus establishing the desired result.

PRACTICE PROBLEMS FOR SECTION 5.3

1. The following data give the number of calls received by a receptionist in 20 randomly selected intervals of one hour each:

16	19	19	15	15	19	17	18	18	20
18	18	17	17	16	16	17	18	18	20

   Find the sample mean $\bar{X}$ and the sample standard deviation S. Find the percentage of data points that fall in each of the following intervals: $(\bar{X} \pm 1.5S)$, $(\bar{X} \pm 2S)$, $(\bar{X} \pm 3S)$. Compare your result with the corresponding percentages given by Chebyshev's inequality.
2. A random variable X is distributed with mean 16 and standard deviation 3. Using Chebyshev's inequality, find lower or upper bounds for the following probabilities:
   (a) $P(|X - 16| \leq 6)$
   (b) $P(|X - 16| \geq 7.5)$
   (c) $P(11.5 \leq X \leq 20.5)$
3. The following data give the number of defective parts manufactured in the last 20 shifts:

6	18	10	12	11	10	11	9	14	7
13	6	9	10	7	12	8	16	12	17

Find the sample mean $\bar{X}$ and the sample standard deviation S. Find the percentage of data points that fall in each of the following intervals: $(\bar{X} \pm 1.5S)$, $(\bar{X} \pm 2S)$, $(\bar{X} \pm 3S)$. Compare your result with the corresponding percentages given by Chebyshev's inequality.

4. According to Chebyshev's inequality, what can we say about the proportion of the data of a given data set that must fall within k standard deviations of the mean, for values of k as follows? (a) $k = 3$, (b) $k = 4$, (c) $k = 8$.
5. Hospital records indicate that patients with heart surgery spend time (in days) in the hospital having a probability distribution with mean six days and standard deviation 1.5 days. Use Chebyshev's inequality to find the lower bound of the percentage of patients who stay between three and nine days (inclusive).
6. Suppose that X is a random variable having probability distribution with mean 16 and standard deviation 2.5. What can you say about the probability $P(8.5 \leq X \leq 23.5)$? (Hint: use Chebyshev's inequality.)

5.4 THE UNIFORM DISTRIBUTION

5.4.1 Definition and Properties

The uniform distribution, because of its shape, is also sometimes known as the rectangular distribution. Because of the shape of its p.d.f., it is perhaps the simplest continuous probability distribution.

Definition 5.4.1 A random variable X is said to be uniformly distributed over an interval $[a, b]$ if its p.d.f. is given by

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for} \quad a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{5.4.1}$$

Note that the density function $f(x)$ in (5.4.1) is constant for all values of x in the interval $[a, b]$. Figure 5.4.1 shows the graphical representation of a uniform distribution of the random variable X distributed over the interval $[a, b]$, where $a < b$.

The probability that the random variable X takes the values in an interval $[x_1, x_2]$, where $a \leq x_1 < x_2 \leq b$ is indicated by the shaded area in Figure 5.4.2 and is equal to

$$P(x_1 \leq X \leq x_2) = \frac{x_2 - x_1}{b - a} \tag{5.4.2}$$

Example 5.4.1 (Uniform distribution) *Let a random variable* X *be the time taken by a technician to complete a project. In this example, assume the time taken by the technician can be anywhere between two to six months and that the random variable* X *is uniformly distributed over the interval* $[2, 6]$. *Find the following probabilities:*
(a) $P(3 \leq X \leq 5)$, *(b)* $P(X \leq 4)$, *and (c)* $P(X \geq 5)$.

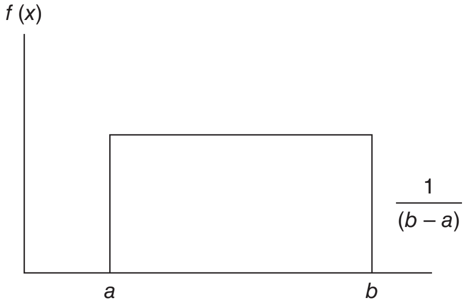


Figure 5.4.1 Uniform distribution over the interval $[a, b]$.

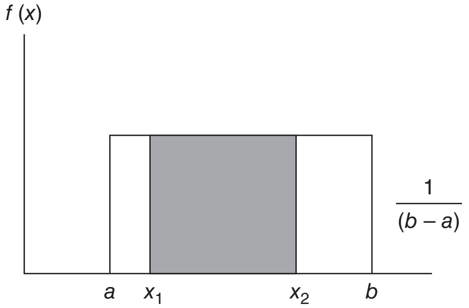


Figure 5.4.2 The probability $P(x_1 \leq X \leq x_2)$ for the uniform random variable X in $[a, b]$.

Solution: (a) To find the probability $P(3 \leq X \leq 5)$, we use the result given in (5.4.2), where we have $a = 2, b = 6, x_1 = 3$, and $x_2 = 5$. Thus,

$$P(3 \leq X \leq 5) = \frac{5-3}{6-2} = \frac{2}{4} = 0.5$$

(b) In this part, we want to find the probability $P(X \leq 4)$. This probability is equivalent to finding the probability $P(2 \leq X \leq 4)$, since the probability for any interval that falls below the point $x = 2$ is zero. Again, using the result in (5.4.2), we have

$$P(X \leq 4) = P(2 \leq X \leq 4) = \frac{4-2}{6-2} = \frac{2}{4} = 0.5$$

(c) By a similar argument as in part (b), $P(X \geq 5) = P(5 \leq X \leq 6)$. Thus,

$$P(X \geq 5) = P(5 \leq X \leq 6) = \frac{6-5}{6-2} = \frac{1}{4} = 0.25$$

Example 5.4.2 (Uniform distribution) *Suppose a delay in starting production due to an unexpected mechanical failure is anywhere from 0 to 30 minutes. Find the following probabilities:*

(a) Production will be delayed by less than 10 minutes.
(b) Production will be delayed by more than 20 minutes.
(c) Production will be delayed by 12–22 minutes.

Solution: Let X be a random variable denoting the time by which production will be delayed. From the given information, we see that the random variable X is uniformly distributed over the interval $[0, 30]$. With this information, the desired probabilities are as follows:

(a) $P(X \leq 10) = P(0 \leq X \leq 10) = \dfrac{10-0}{30-0} = \dfrac{10}{30} = \dfrac{1}{3}$

(b) $P(X > 20) = P(20 < X \leq 30) = \dfrac{30-20}{30-0} = \dfrac{10}{30} = \dfrac{1}{3}$

(c) $P(12 \leq X \leq 22) = \dfrac{22-12}{30-0} = \dfrac{10}{30} = \dfrac{1}{3}$

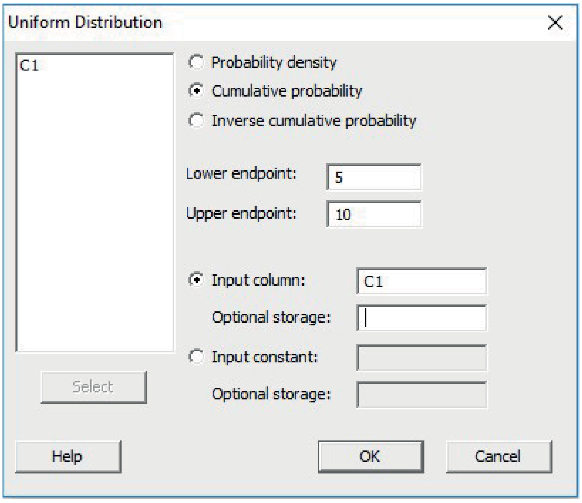
In each case of this example, the probability turned out to be 1/3, demonstrating that in a uniform distribution, the probability depends upon the length of the interval and not on the location of the interval. In each case, the length of the interval is equal to 10.

Example 5.4.3 (Uniform distribution) *A random variable* X *is the time taken by an engineer to develop a design of a new product. Here, the time taken by the engineer is anywhere between five to ten months, so that the random variable* X *is uniformly distributed over the interval* $[5, 10]$*. Find the probability* $P(6 \leq X \leq 8)$*, using MINITAB and R.*

Solution: In order to determine the probability $P(6 \leq X \leq 8)$, we first have to find the probabilities $P(X \leq 6)$ and $P(X \leq 8)$. Then, $P(6 \leq X \leq 8) = P(X \leq 8) - P(X \leq 6)$. To find the probabilities $P(X \leq 8)$ and $P(X \leq 6)$, we proceed as follows:

MINITAB

1. Enter the values 6 and 8 in column C1.
2. From the menu bar, select **Calc** > **Probability Distribution** > **Uniform**.
3. In the dialog box that appears on the screen, click the circle next to **Cumulative probability**.


Uniform Distribution

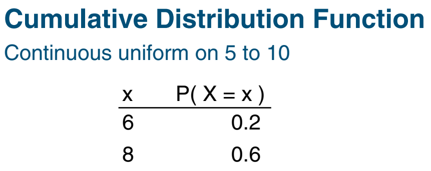
C1

○ Probability density
● Cumulative probability
○ Inverse cumulative probability

Lower endpoint: 5
Upper endpoint: 10

● Input column: C1
Optional storage:
○ Input constant:
Optional storage:

4. Enter 5 in the box next to **Lower endpoint** and 10 in the box next to **Upper endpoint**. Click the circle next to **Input column** and type C1 in the box next to it.
5. Click **OK**.
6. In the session window, text will appear as shown in the following box. Using this, we have that $P(6 \leq X \leq 8) = P(X \leq 8) - P(X \leq 6) = 0.6 - 0.2 = 0.4$.



Cumulative Distribution Function

Continuous uniform on 5 to 10

x	P(X = x)
6	0.2
8	0.6

USING R

R has a built in cumulative uniform distribution function 'punif(q, min, max)', where q is the quantile, and min and max are the lower and upper boundaries of the uniform distribution, respectively. So, referring to Example 5.4.3, we are asked to find $P(6 \leq X \leq 8)$, which can be computed by typing punif(8, 5, 10) - punif(6, 5, 10) in the R Console window as follows.

```
punif(8, 5, 10) - punif(6, 5, 10)

#R output
[1] 0.4
```

5.4.2 Mean and Standard Deviation of the Uniform Distribution

Let X be a random variable distributed uniformly over an interval $[a, b]$. Then, the reader should verify that the *mean* μ and the *variance* σ^2 of the random variable X are given by

$$\mu = E(X) = \int_a^b x\left(\frac{1}{b-a}\right)dx = \frac{a+b}{2} \tag{5.4.3}$$

$$\sigma^2 = E(X-\mu)^2 = \int_a^b (x-\mu)^2\left(\frac{1}{b-a}\right)dx = \frac{(b-a)^2}{12} \tag{5.4.4}$$

The *m.g.f.* of the random variable X having a uniform distribution given in Equation (5.4.1) is (as the reader should verify) stated in (5.4.5).

$$M_X(t) = \begin{cases} \dfrac{e^{tb}-e^{ta}}{t(b-a)}, & t \neq 0 \\ 1, & t = 0 \end{cases} \tag{5.4.5}$$

The *c.d.f.* $F(x)$ of a random variable X distributed uniformly over an interval $[a, b]$ is given by

$$\begin{aligned} F(x) &= P(X \leq x), \quad a \leq x \leq b \\ &= P(a \leq X \leq x) \\ &= \frac{x-a}{b-a}, \quad a \leq x \leq b \end{aligned} \tag{5.4.6}$$

We note that $F(x) = 0$, if $x < a$ and $F(x) = 1$, if $x > b$.

Example 5.4.4 (Uniform distribution) *A random variable* X *denotes the time spent (in minutes) for a coffee break that a technician takes every morning. Suppose this random variable* X *is uniformly distributed over an interval* $[0, 16]$*. Find the mean* μ *and the standard deviation* σ *of the distribution.*

Solution: Using Equations (5.4.3) and (5.4.4), we get

$$\mu = \frac{a+b}{2} = \frac{0+16}{2} = 8, \quad \sigma = \frac{b-a}{\sqrt{12}} = \frac{16-0}{\sqrt{12}} = 4.619$$

Example 5.4.5 (Example 5.4.4, continued) *In Example 5.4.4, find the following values of the distribution function of the random variable* X*: (a)* $F(3)$ *(b)* $F(5)$ *(c)* $F(12)$*.*

Solution: Using the result of Equation (5.4.6), we obtain

(a) $F(3) = \dfrac{x-a}{b-a} = \dfrac{3-0}{16-0} = \dfrac{3}{16}$

(b) $F(5) = \dfrac{x-a}{b-a} = \dfrac{5-0}{16-0} = \dfrac{5}{16}$

(c) $F(12) = \dfrac{x-a}{b-a} = \dfrac{12-0}{16-0} = \dfrac{3}{4}$

PRACTICE PROBLEMS FOR SECTION 5.4

1. The time taken to download software from the internet is uniformly distributed between four and 10 minutes.
   (a) What is the probability that the downloading time for software is more than six minutes?
   (b) What is the probability that the downloading time for software is between five and eight minutes?
   (c) What is the probability that the downloading time for software is less than seven minutes?
2. The city buses stop at an assigned point every 15 minutes. A passenger who takes the bus from that point has arrived there but does not know when the last bus came.
   (a) What is the probability that the passenger will have to wait for more than eight minutes to take the next bus?

(b) What is the probability that the passenger will have to wait for less than 10 minutes to take the next bus?
(c) What is the probability that the passenger will have to wait for five to nine minutes to take the next bus?

3. Referring to Problem 2, let the random variable X denote the time that the passenger has to wait. Find the mean, variance, and standard deviation of the random variable X.
4. A random variable X is uniformly distributed with mean 12 and variance three. Find the m.g.f. of the random variable X.
5. The hourly wages of certain group of workers in a large manufacturing company are uniformly distributed on the interval $[20, 32]$. What percentage of these workers are making over \$25 an hour?
6. Suppose that a random variable X is distributed uniformly over an interval $[15, 25]$. Find the following probabilities: (a) $P(20 \leq X \leq 25)$, (b) $P(X \leq 25)$, (c) $P(15 \leq X \leq 25)$, (d) $P(X \leq 15)$.
7. The time X (in hours) taken by students to complete a standardized test is uniformly distributed over the interval $[2, 4]$. Find the mean and the standard deviation of X. Then, find the probability $P(\mu - 2\sigma < X < \mu + 2\sigma)$.
8. A manufacturing company has designed a fuel-injection system for medium-size cars such that the cars yield X mpg uniformly distributed over an interval [45, 50]. Find the mean and variance of X. Find the probability $P(X > \mu - \sigma)$.

5.5 THE NORMAL DISTRIBUTION

5.5.1 Definition and Properties

The normal distribution plays a fundamental role in all of mathematical statistics, and, as we shall see in later chapters, important statistical techniques are based on this distribution. The purpose of this section is to discuss the normal distribution, its properties, and some of its applications.

Definition 5.5.1 A random variable X is said to be distributed by the *normal distribution*, say $N(\mu, \sigma^2)$, if its p.d.f. is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], -\infty < x < +\infty \tag{5.5.1}$$

where μ and σ are parameters that obey the conditions $-\infty < \mu < +\infty$ and $\sigma > 0$.

We shall show that the parameters μ and σ in the above definition are the *mean* and *standard deviation* of X, respectively. We note that $f(x) > 0$ for $-\infty < x < +\infty$ As may be proved, the density (5.5.1) is such that $\int_{-\infty}^{\infty} f(x)dx = 1$. The graph of the p.d.f. $f(x)$ given by (5.5.1) is shown in Figure 5.5.1.

To determine the mean of X, we have

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] dx \tag{5.5.2}$$

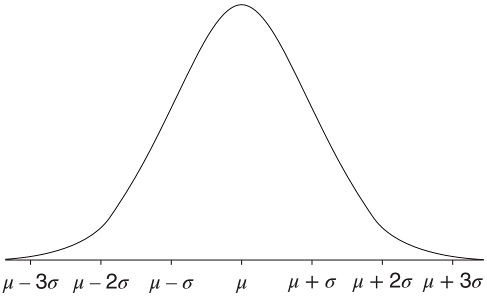


Figure 5.5.1 Graph of the normal p.d.f. given in (5.5.1).

In the integral, set $(x-\mu)/\sigma = z$. Then (5.5.2) takes the form

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}(\mu+\sigma z)e^{-z^2/2}dz$$

or

$$E(X) = \mu\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz + \sigma\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}ze^{-z^2/2}dz$$

The integrand of the first integral in the preceding line is that of the normal density function (5.5.1) having $\mu = 0, \sigma = 1$. Hence, the value of this integral is 1. For the second integral, we have

$$\sigma\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}ze^{-z^2/2}dz = \left[-\frac{e^{-z^2/2}}{\sqrt{2\pi}}\right]_{-\infty}^{\infty} = 0$$

Putting these results together, we find that the *mean* or the *expected value* $E(X)$ of the normal random variable X with p.d.f. given by (5.5.1) is μ. Similarly, it can be shown that for the variance of X, we have

$$Var(X) = E[X-E(X)]^2 = E(X-\mu)^2 = \sigma^2 \tag{5.5.3}$$

That is, we have that the *variance* of the normal random variable X is σ^2. In summary, we find that the p.d.f. $f(x)$ of the normal distribution $N(\mu, \sigma^2)$ given by (5.5.1) has its mean μ and standard deviation σ built in as parameters.

Some of the characteristics of the normal density function are listed below.

1. The normal density function curve is bell-shaped and completely symmetric about its mean μ. For this reason, the normal distribution is also known as a bell-shaped distribution.
2. The tails of the density function extend from $-\infty$ to $+\infty$.
3. The total area under the curve is 1.0. However, 99.73% of the area falls within three standard deviations of the mean μ.
4. The area under the normal curve to the right of μ is 0.5 and to the left of μ is also 0.5.
5. As the mean μ and the standard deviation change, the location and the shape of the normal curve change (see Figures 5.5.2 and 5.5.3).

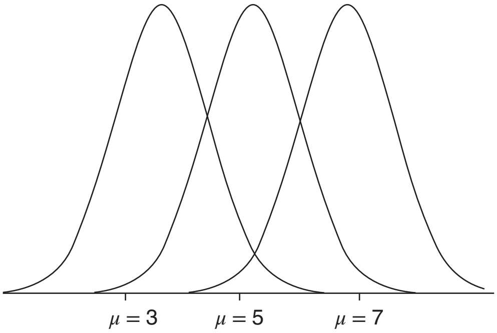


Figure 5.5.2 Curves representing the normal density function with different means but the same standard deviation.

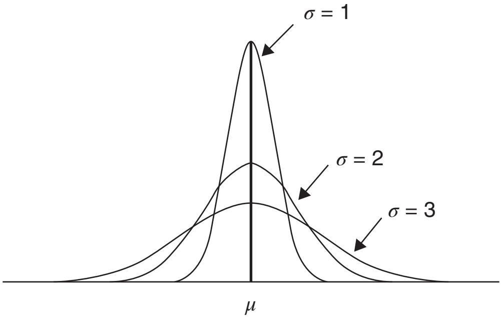


Figure 5.5.3 Curves representing the normal density function with different standard deviations but the same mean.

5.5.2 The Standard Normal Distribution

A random variable having the standard normal distribution $N(0,1)$ has the special designation Z throughout this book, and its particular values are designated by z. We then call Z the *standard normal variable*, and its p.d.f. is usually denoted by $f(z)$, where

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < +\infty \tag{5.5.4}$$

By comparing (5.5.1) with (5.5.4), it is evident that $E(Z) = 0$ and $Var(Z) = 1$. To find values of the c.d.f. of Z, denoted by $\Phi(z)$, we use Table A.4, which provides values of $P(Z \leq z) = \Phi(z)$ for all z such that $(-3.4 \leq z \leq 3.4)$ in increments of 0.01. The c.d.f. of Z is given below.

$$\Phi(z) = \int_{-\infty}^{z} f(x)dx \tag{5.5.5}$$

In view of the symmetry of $f(z)$ about $z = 0$, the reader should note that, for any z,

$$\Phi(-z) = 1 - \Phi(z) \tag{5.5.6}$$

We may restate (5.5.6) as

$$P(Z \le -z) = 1 - P(Z \le z) = P(Z \ge z)$$

that is,

$$P(Z \le -z) = P(Z \ge z)$$

By using tables of $\Phi(z)$, we can find the probabilities associated with events concerning any normal random variable. Suppose we have a random variable X that has the distribution $N(\mu, \sigma^2)$, and, for any given value x', we wish to find the probability

$$P(X \le x') = \int_{-\infty}^{x'} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \tag{5.5.7}$$

This probability is represented by the shaded area in Figure 5.5.4. We do not have tables of $P(X \le x')$ for every possible value of μ and σ^2 because we do not need them. Suppose in the integral (5.5.7) we let $z = (x - \mu)/\sigma$. Then, it easily follows that

$$P(X \le x') = \int_{-\infty}^{(x'-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi\left(\frac{x'-\mu}{\sigma}\right) = P\left(Z \le \frac{x'-\mu}{\sigma}\right) \tag{5.5.8}$$

where Z is the standard normal variable. Thus, in order to evaluate $P(X \le x')$, where X has the normal distribution $N(\mu, \sigma^2)$, we note that the event $(X \le x')$ is equivalent to the event $(X - \mu \le x' - \mu)$, which in turn is equivalent to the event $\left(\frac{X-\mu}{\sigma} \le \frac{x'-\mu}{\sigma}\right)$. Hence, we write

$$P(X \le x') = P(X - \mu \le x' - \mu) = P\left(\frac{X-\mu}{\sigma} \le \frac{x'-\mu}{\sigma}\right) \tag{5.5.9}$$

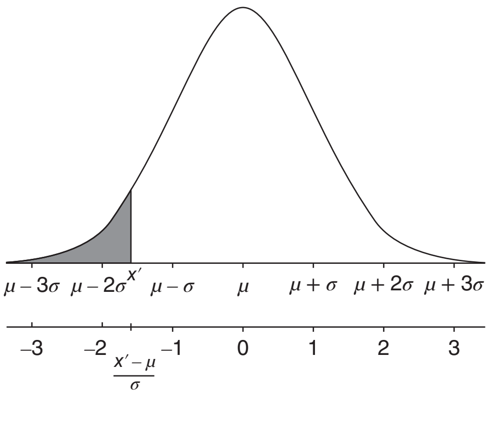


Figure 5.5.4 Graph of the normal p.d.f. $f(x)$ given by (5.5.1) showing the relationship between the x-axis and z-axis. The shaded area of this normal density is $P(X \le x')$ as given by (5.5.7).

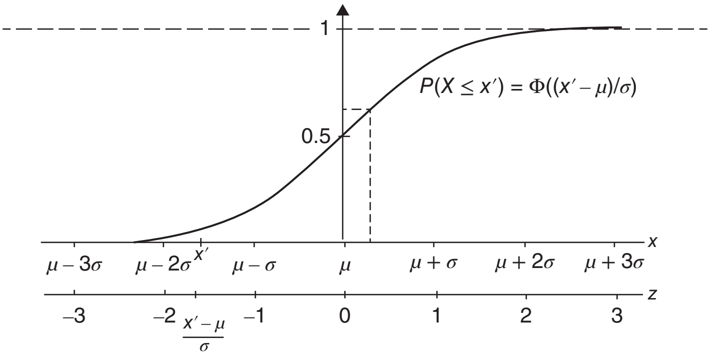


Figure 5.5.5 Graph of the normal c.d.f. (5.5.7), showing the relationship between the x-axis and the z-axis.

But from (5.5.8), we have

$$P\left(\frac{X-\mu}{\sigma} \le \frac{x'-\mu}{\sigma}\right) = P\left(Z \le \frac{x'-\mu}{\sigma}\right) = \Phi\left(\frac{x'-\mu}{\sigma}\right) \tag{5.5.10}$$

That is, the probability that $X \le x'$ is the same as the probability that $P(Z \le (x'-\mu)/\sigma)$, where Z is the standardized normal variable having a p.d.f. $f(z)$ given by (5.5.4). The graph of the c.d.f. $P(X \le x')$ as given by (5.5.7) is shown in Figure 5.5.5. The relationship between the x-scale and the z-scale is shown graphically in Figures 5.5.4 and 5.5.5.

For convenience, we often abbreviate the phrase "X is a random variable having the normal distribution $N(\mu, \sigma^2)$" by saying "X is a $N(\mu, \sigma^2)$ random variable" or "X is from a $N(\mu, \sigma^2)$ population."

Example 5.5.1 (Normally distributed quality characteristic) *Suppose a quality characteristic of a product is normally distributed with mean* $\mu = 18$ *and standard deviation* $\sigma = 1.5$*. The specification limits furnished by the customer are* $(15, 21)$*. Determine what percentage of the product meets the specifications set by the customer.*

Solution: The random variable X denotes the quality characteristic of interest. Then, X is normally distributed with mean $\mu = 18$ and standard deviation $\sigma = 1.5$. We are interested in finding the percentage of product with the characteristic of interest within the limits $(15, 21)$, which is equivalent to finding the probability $P(15 \le X \le 21)$ and multiplying it by 100. That is, first we determine

$$\begin{aligned}
P(15 \le X \le 21) &= P\left(\frac{15-18}{1.5} \le \frac{X-18}{1.5} \le \frac{21-18}{1.5}\right)\\
&= P(-2.0 \le Z \le 2.0)\\
&= P(Z \le 2.0) - P(Z \le -2.0)\\
&= 0.9772 - 0.0228\\
&= 0.9554
\end{aligned}$$

Then, the percentage of product that will meet the specifications set by the customer is 95.44%.

Example 5.5.2 (Normally distributed pin-diameters) *A manufacturer knows from experience that the diameters of 0.250 in. precision-made pins he produces have a normal distribution with mean 0.25000 in. and standard deviation 0.00025 in. What percentages of the pins have diameters between 0.24951 and 0.25049 in? This question is equivalent to finding the probability that the diameter, say X, of a pin taken at random from the production lies between 0.24951 and 0.25049 in.*

Solution: We must find $P(0.24951 \le X \le 0.25049)$, where X has the distribution $N(0.25000, (0.00025)^2)$. We proceed with operations indicated in (5.5.10) and obtain:

$$\begin{aligned}
P(0.24951 \le X \le 0.25049) &= P\left(\frac{0.24951 - 0.25000}{0.00025} \le \frac{X - 0.25000}{0.00025} \le \frac{0.25049 - 0.25000}{0.00025}\right)\\
&= P(-1.96 \le Z \le 1.96)\\
&= P(Z \le 1.96) - P(Z \le -1.96)\\
&= 0.975 - 0.025 = 0.95
\end{aligned}$$

That is, 95% of the production lies between 0.24951 and 0.25049 in.

Example 5.5.3 (Production process of chalks) *A process for producing batches of chalk is such that the bulk density* X *of a batch of chalk is a normally distributed random variable with mean 0.8000g/cc and standard deviation 0.0030 g/cc. Find: (a)* $P(X \le 0.8036)$*, (b)* $P(|X - 0.8000| \le 0.0060)$*. (c) Find* c *such that* $P(X \le c) = 0.95$

Solution: Here, X is an $N(0.8000, (0.003)^2)$ variable. Thus,

(a)
$$\begin{aligned}
P(X \le 0.8036) &= P\left(\frac{X - 0.8000}{0.0030} \le \frac{0.8036 - 0.8000}{0.0030}\right)\\
&= P(Z \le 1.2) = 0.8849
\end{aligned}$$

(b)
$$\begin{aligned}
P(|X - 0.8000| \le 0.0060) &= P\left(\frac{|X - 0.8000|}{0.0030} \le \frac{0.0060}{0.0030}\right)\\
&= P(|Z| \le 2) = P(-2 \le Z \le 2)\\
&= P(Z \le 2) - P(Z \le -2)\\
&= 0.97725 - 0.02275 = 0.95450
\end{aligned}$$

(c) From $P(X \le c) = 0.95$, we have that

$$P\left(\frac{X - 0.8000}{0.0030} \le \frac{c - 0.8000}{0.0030}\right) = P\left(Z \le \frac{c - 0.8000}{0.0030}\right) = 0.95$$

This implies that $\dfrac{c - 0.8000}{0.0030} = 1.645$, since $P(Z \leq 1.645) = \Phi(1.645) = 0.95$ (see Table A.4). Thus, we have

$$c = 0.8000 + (0.0030)(1.645) = 0.804935$$

Throughout this book, we use the notation z_α to denote the $100(1 - \alpha)$ *percentage point* of the standard normal distribution, so that z_α is determined by:

$$P(Z \geq z_\alpha) = \int_{z_\alpha}^{\infty} f(z)dz = \alpha$$

or expressed alternatively,

$$\Phi(z_\alpha) = P(Z \leq z_\alpha) = 1 - \alpha$$

Note that $z_\alpha = -z_{1-\alpha}$. For convenience, we list a few of the most important and commonly used values of z_α in Table 5.5.1.

Table 5.5.1 Some percentage points of the normal distribution.

α	0.10	0.05	0.025	0.01	0.005
z_α	1.282	1.645	1.960	2.326	2.576

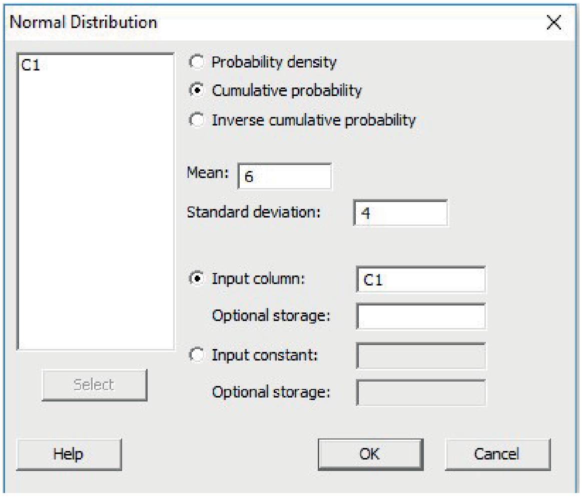
We often note that z_α is the point exceeded with probability α when using the standard normal distribution. The percentage points z_α are often called *significance points.* For example, $z_{0.05} = 1.645$ is the 5% significance point of the standard normal distribution.

Example 5.5.4 (Using MINITAB and R for normal probabilities) *A random variable* X *is distributed as normal with mean $\mu = 6$ and standard deviation $\sigma = 4$. Determine the probability $P(8.0 \leq X \leq 14.0)$, using MINITAB and R.*

Solution: In order to determine the probability $P(8.0 \leq X \leq 14.0)$, we first have to find the probabilities $P(X \leq 8.0)$ and $P(X \leq 14.0)$. Then, $P(8.0 \leq X \leq 14.0) = P(X \leq 14.0) - P(X \leq 8.0)$. To find the probabilities $P(X \leq 8.0)$ and $P(X \leq 14.0)$, we proceed as follows:

MINITAB

1. Enter the values 8 and 14 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **Normal**.
3. In the dialog box, click the circle next to **Cumulative probability**.
4. Enter 6 (the value of the mean) in the box next to **Mean** and 4 (the value of the standard deviation) in the box next to **Standard deviation**.
5. Click the circle next to **Input column** and type C1 in the box next to it.
6. Click **OK**.



Normal Distribution

C1

○ Probability density
● Cumulative probability
○ Inverse cumulative probability

Mean: 6

Standard deviation: 4

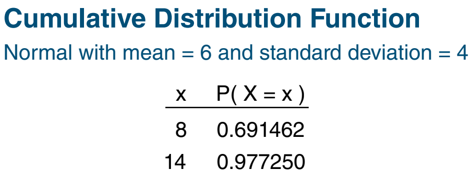
● Input column: C1

Optional storage:

○ Input constant:

Optional storage:

7. In the session window, text will appear as shown in the following box. Thus, $P(8.0 \leq X \leq 14.0) = P(X \leq 14.0) - P(X \leq 8.0) = 0.977250 - 0.691462 \approx 0.2858$



Cumulative Distribution Function

Normal with mean = 6 and standard deviation = 4

x	P(X = x)
8	0.691462
14	0.977250

USING R

R has a built in cumulative normal distribution function 'pnorm(q, mean, sd)', where q is the quantile, and mean and sd are the mean and the standard deviation of the normal distribution, respectively.

So, referring to Example 5.5.4, we are asked to find $P(8.0 \leq X \leq 14.0)$, which can be computed by typing the following in the R Console window.

```
pnorm(14, 6, 4) - pnorm(8, 6, 4)

#R output
[1] 0.2857874
```

MINITAB and R answers are equal up to the fifth decimal place.

5.5.3 The Moment-Generating Function of the Normal Distribution

From our discussion of the m.g.f. in Chapter 4 and Equation (5.2.12), it follows that the m.g.f. of a random variable X having the normal distribution $N(\mu, \sigma^2)$ is: (The derivation is not given here but is available on the book website at: www.wiley.com/college/gupta/statistics2e)

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2) \tag{5.5.11}$$

Differentiating $M_X(t)$ with respect to t, we have

$$M'_X(t) = (\mu + \sigma^2 t)e^{\mu t + \sigma^2 t^2/2}$$
$$M''_X(t) = \sigma^2 e^{\mu t + \sigma^2 t^2/2} + (\mu + \sigma^2 t)^2 e^{\mu t + \sigma^2 t^2/2}$$

and hence $M'_X(0) = \mu$, $M''_X(0) = \sigma^2 + \mu^2$, that is, $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$. We thus have confirmation of (5.5.3) that

$$E(X) = \mu$$

$$Var(X) = E(X^2) - [E(X)]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

Note that the m.g.f. of the normal distribution has, in the exponent of e, its mean as the coefficient of t, and the variance of the distribution as the coefficient of $(1/2)t^2$. Summarizing, we may state the following theorem.

Theorem 5.5.1 *If X is an $N(\mu, \sigma^2)$ random variable, then the m.g.f. $M_X(t)$ is given by*

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

where $E(X) = \mu$ and $Var(X) = \sigma^2$.

PRACTICE PROBLEMS FOR SECTION 5.5

1. A random variable X is normally distributed with mean $\mu = 10$ and standard deviation $\sigma = 1.5$. Find the following probabilities: (a) $P(8 \leq X \leq 12)$, (b) $P(X \leq 12)$, (c) $P(X \geq 8.5)$.
2. A random variable X is normally distributed with unknown mean μ and unknown standard deviation σ. Find μ and σ if it is known that the probability that X is less than 10 is 0.6950 and the probability that X exceeds 6 is 0.7939.
3. The weights of Maine lobsters at the time of their catch are normally distributed with a mean of 1.8 lb and a standard deviation of 0.25 lb. What is the probability that a randomly selected lobster weighs (a) between 1.5 and 2 lb?, (b) more than 1.55 lb?, (c) less than 2.2 lb?
4. The postsurgery survival time of a breast cancer patient is normally distributed with a mean of eight years and a standard deviation of 1.5 years. Find the probabilities that a woman with breast cancer will survive after her surgery: (a) between five and seven years, (b) more than 11 years, (c) less than six years.
5. The amount of beverage in a 16-oz bottle is normally distributed with a mean of 16.2 oz and a standard deviation of 0.1 oz. Find the probability that the amount of beverage in a randomly select can is (a) between 15.5 and 16.2 oz, (b) more than 16.4 oz, (c) less than 16.1 oz.

6. The total cholesterol level (LDL + HDL + 20% of triglyceride) of US males between 60 and 70 years of age is approximately normally distributed with a mean of 175 mg/100 ml and a standard deviation of 20 mg/100 ml. Find the probability that a randomly selected person from this population has total cholesterol level: (a) between 150 and 250 mg/100 ml, (b) more than 155 mg/100 ml, and (c) less than 215 mg/100 ml.
7. The height of a certain population of female teenagers is approximately normally distributed with a mean of 155 cm and a standard deviation of 7 cm. Find the probability that a randomly selected teenager has height (a) between 145 and 165 cm, (b) more than 150 cm, and (c) less than 169 cm.
8. A car manufacturer claims that its new hybrid car can travel at speeds more than 60 mpg. The actual mpg is approximately normally distributed with a mean of 52 miles and a standard deviation of three miles. What is the probability that the manufacturer's claim is valid? Comment on this claim.

5.6 DISTRIBUTION OF LINEAR COMBINATION OF INDEPENDENT NORMAL VARIABLES

In general, we know that if $X_1, \ldots, X_n$ are independent random variables with means and variances $\mu_i, \sigma_i^2,\ i = 1, 2, \ldots, n$, respectively, then a linear combination of the X_i, say

$$Y = \sum_{i=1}^{n} c_i X_i$$

is such that

$$\mu_Y = E(Y) = \sum_{i=1}^{n} c_i \mu_i \tag{5.6.1}$$

and

$$\sigma_Y^2 = E(Y - \mu_Y)^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2 \tag{5.6.2}$$

Indeed, if $X_1, \ldots, X_n$ are normal independent random variables, then we have an important result about the normal distribution that may be stated as in Theorem 5.6.1.

Theorem 5.6.1 *Let $X_1, \ldots, X_n$ be* n *random independent variables having distribution $N(\mu_1, \sigma_1^2), \ldots, N(\mu_n, \sigma_n^2)$. Then, the random variable $Y = \sum_{i=1}^{n} c_i X_i$, that is, a linear combination of independent normal variables, is normally distributed with mean $\sum_{i=1}^{n} c_i \mu_i$ and variance $\sum_{i=1}^{n} c_i^2 \sigma_i^2$.*

The proof of this theorem follows immediately using the result of Theorem 5.5.1, and the fact that the m.g.f. of the sum of independent random variables is the product of their m.g.f.s.

Now, if the X_i's are $N(\mu, \sigma^2)$ random variables and are independent, that is, if $(X_1, \ldots, X_n)$ is a *random sample of size* n from $N(\mu, \sigma^2)$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{1}{n} X_i$$

is a linear combination of the X_i with $c_i = 1/n,\ i = 1, 2, \ldots, n$.

Hence, we have that the sample mean or average, $\bar{X}$, is itself normally distributed with mean and variance given by

$$\mu_{\bar{x}} = \sum_{i=1}^{n} c_i \mu_i = \sum_{i=1}^{n} \frac{1}{n}\mu = \mu \tag{5.6.3}$$

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2 = \sum_{i=1}^{n} \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n} \tag{5.6.4}$$

This is a very important result that we restate in the following theorem.

Theorem 5.6.2 *If $X_1, \ldots, X_n$ is a random sample of size n from $N(\mu, \sigma^2)$, then $\bar{X}$ is a $N(\mu, \sigma^2/n)$ random variable.*

We note that if T is the sample sum defined by

$$T = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i \tag{5.6.5}$$

then T may be regarded as linear combination of the X_i with $c_i = 1,\ i = 1, 2, \ldots, n$. Thus, we have the following theorem, whose proof follows from Theorem 5.6.1.

Theorem 5.6.3 *If $X_1, \ldots, X_n$ is a random sample of size* n *from $N(\mu, \sigma^2)$, then $T = \sum_{i=1}^{n} X_i$ is a $N(n\mu, n\sigma^2)$ random variable.*

If the sample $X_1, \ldots, X_n$ is a random sample from any population having mean μ and variance σ^2 (both finite), it can be shown that as $n \to \infty$, $(T - n\mu)/\sqrt{n\sigma^2}$ or equivalently $(\bar{X} - \mu)\sqrt{n}/\sigma$ is a random variable having $N(0, 1)$ as its limiting distribution. This result is known as the *central limit theorem.* We note that this theorem has the implication that even though $X_i, i = 1, \ldots, n$, is a random sample of size n from a nonnormal distribution, if n is large, T is approximately an $N(n\mu, n\sigma^2)$ random variable, with error of approximation tending to zero as $n \to \infty$. We will revisit this theorem in Chapter 7 on sampling distributions.

In engineering statistics, there often arises a need to compare two populations or processes involving two independent random variables X and Y, and in many cases, it is useful to examine the difference (or the contrast) $L = X - Y$. If we know the means and variances of X and Y, we can compute the mean and variance of L from (5.6.1) and (5.6.2) to find

$$\mu_L = E(L) = \mu_x - \mu_y \tag{5.6.6}$$

$$\sigma_L^2 = Var(L) = \sigma_x^2 + \sigma_y^2 \tag{5.6.7}$$

and hence the standard deviation of L is given by

$$\sigma_L = Var(L) = \sqrt{\sigma_x^2 + \sigma_y^2} \tag{5.6.8}$$

Furthermore, it can easily be shown that if X and Y are normally distributed, so is L.

Example 5.6.1 (Matching normally distributed pin diameters) *In manufacturing precision-made pins, assume that "0.5-inch pins" have diameters that are (approximately) normally distributed with mean 0.4995 in. and standard deviation 0.0003 in. and that matching parts with holes to receive these pins have diameters that are (approximately) normally distributed with mean 0.5005 and standard deviation 0.0004 in. If pins and holes are matched at random, in what fraction of the matches would the pin fit?*

Solution: Let D_p be the diameter in inches of a pin and D_m be the diameter in inches of the hole of a matching part. Consider $D = D_m - D_p$. We then have a fit if $D > 0$. Now, the mean of D is

$$\mu_d = \mu_{d_m} - \mu_{d_p} = 0.5005 - 0.4995 = 0.0010$$

and the variance of D is given by

$$\sigma_d^2 = \sigma_{d_m}^2 + \sigma_{d_p}^2 = 16 \times 10^{-8} + 9 \times 10^{-8} = 25 \times 10^{-8}$$

so that the standard deviation of D is $\sigma_d = 5 \times 10^{-4} = 0.0005$. Furthermore, D is a linear combination of normally distributed variables and thus is itself normally distributed. Again, a pin fits a matching part if, and only if, $D > 0$, and the probability of this is

$$\begin{aligned} P(D > 0) &= P\left(\frac{D - 0.0010}{0.0005} > \frac{-0.0010}{0.0005}\right) \\ &= P(Z > -2) = 0.9772 \end{aligned}$$

That is, in 97.72% of the matches, the pins would fit.

Example 5.6.2 (Boxes filled with corn flakes) *The distribution of gross weights of 8-oz boxes of cornflakes is known to be normal with mean 9.60 oz and standard deviation 0.80 oz. Suppose that the boxes are packed 24 to a carton and the population of weights of empty cartons is also normal with mean 24.00 oz and standard deviation 2.20 oz. Determine the mean and the variance of the population of weights of filled cartons. What percentage of the filled cartons will have weights between 250 and 260 oz?*

Solution: Let X_i be the gross weight of the ith box in the sample of 24 boxes, $i = 1, 2, \ldots, 24$, and let T be the total weight of 24 boxes of cornflakes, that is,

$$T = X_1 + X_2 + \cdots + X_{24}$$

From (5.6.1) and (5.6.2) we have, since $\mu_{x_i} = \mu_x = 9.60$, that

$$\mu_T = \mu_x + \mu_x + \cdots + \mu_x = 24\mu_x = 24(9.60) = 230.4 \text{ oz}$$

and since $\sigma^2_{x_i} = \sigma^2_x = (0.8)^2 = 0.64$, σ^2_T is given by

$$\sigma^2_T = \sigma^2_x + \sigma^2_x + \cdots + \sigma^2_x = 24\sigma^2_x = 24(0.64) = 15.36 \text{ oz}^2$$

Let $W = T + Y$, where Y is the weight of an empty carton. Then,

$$\begin{aligned}
\mu_W &= \mu_T + \mu_Y = 230.4 + 24.0 = 254.4 \text{ oz}\\
\sigma^2_W &= \sigma^2_T + \sigma^2_Y = 15.36 + 4.84 = 20.2 \text{ oz}^2\\
\sigma_W &= 4.494 \text{ oz}
\end{aligned}$$

To answer the second question, we must evaluate $P(250 < W < 260)$, which, when reduced in terms of the standard normal variable Z, has the value

$$P(-0.979 < Z < 1.246) = \Phi(1.246) - \Phi(-0.979) = \Phi(1.246) - (1 - \Phi(0.979)) = 0.7299.$$

Thus, approximately 73% of the filled cartons have weights between 250 and 260 oz.

Often, situations arise that uses the result of the useful theorem 5.6.4 given below.

Theorem 5.6.4 *Let $\bar{X}_1$ and $\bar{X}_2$ be sample means of independent samples of sizes n_1 and n_2 from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Then, $\bar{X}_1 - \bar{X}_2$ is distributed normally with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. The proof of this theorem follows directly from Equations (5.6.3), (5.6.4), and Theorem 5.6.2 along with Equations (5.6.6) and (5.6.7).*

PRACTICE PROBLEMS FOR SECTION 5.6

1. Suppose that independent random variables, say X and Y, are normally distributed with means of 10 and 15, and standard deviations of 3 and 4, respectively. Find the following probabilities: (a) $P(X + Y \geq 33)$, (b) $P(-8 \leq X - Y \leq 6)$, (c) $P(20 \leq X + Y \leq 28)$, (d) $P(X - 2Y \leq -10)$.
2. Suppose that independent random variables X and Y are normally distributed with means of 12 and 15 and standard deviations of 4 and 5, respectively. Find the m.g.f. of the random variable $X + 2Y$.
3. The times required to finish two projects, say X and Y, are independently and normally distributed with means of 70 and 75 minutes and standard deviations of 8 and 10 minutes, respectively. Find the following probabilities: (a) $P(X + Y \geq 145)$, (b) $P(-18 \leq X - Y \leq 16)$, (c) $P(122 \leq X + Y \leq 168)$.
4. Referring to Problem 3, determine the probability distribution of the random variable $U = 2X + 3Y$.

5. Scores obtained by students in three sections of the Medical College Admission Test (MCAT) are independently normally distributed with means 10, 12, and 10 and standard deviations 2.6, 1.2, and 1.3, respectively. Determine the probability distribution of the total scores obtained by these students.
6. Suppose in Problem 5 the total scores of a student are denoted by a random variable U. Then, find the following probabilities: (a) $P(U \geq 33)$, (b) $P(30 \leq U \leq 38)$, (c) $P(U \geq 38)$.

5.7 APPROXIMATION OF THE BINOMIAL AND POISSON DISTRIBUTIONS BY THE NORMAL DISTRIBUTION

5.7.1 Approximation of the Binomial Distribution by the Normal Distribution

We now turn to a remarkable application of the normal distribution, namely, that of approximating probabilities associated with the binomial distribution. We recall from Chapter 4 that if X is a random variable having the binomial distribution ($X \sim b(x|n;p)$), then its mean is np, and its variance is $np(1-p)$. We now state a very useful theorem:

Theorem 5.7.1 *As $n \to \infty$, the distribution of the random variable*

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \tag{5.7.1}$$

has the $N(0,1)$ distribution as its limiting distribution.

This theorem was first proved by the French mathematician Abraham de Moivre. Use of this theorem enables us to approximate, for large n, sums of probabilities given by the binomial distribution by appropriate integrals of the standard normal distribution. More precisely, the theorem states that for large n, the random variable X has approximately the normal distribution $N(np, np(1-p))$. The approximation improves as n increases and is "quite good" for values of "p" not too close to zero or one. Further the approximation is appropriate when n and p are such that $np > 5$ and $n(1-p) > 5$.

Example 5.7.1 (Binomial probabilities using the normal approximation) *If* X *is a random variable having the binomial probability function*

$$b(x) = \binom{16}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{16-x}, \quad x = 0, 1, 2, \ldots, 16$$

approximate the value of $P_B(6 \leq X \leq 10)$ using the normal distribution. Here, $n = 16$ and $p = 1/2$, so that $np = 8 > 5$ and $n(1-p) = 8 > 5$. The exact value of the *required probability is*

$$P_B(6 \leq X \leq 10) = \sum_{x=6}^{10} \binom{16}{x} \left(\frac{1}{2}\right)^{16}$$

By using the normal approximation, say $N(\mu, \sigma^2)$, we set $\mu = E(X) = np = 8, \sigma^2 = Var(X) = np(1-p) = 4$, and we find that after matching a so-called "half-integer" or "continuity" correction, (see Figure 5.7.1), that

$$\begin{aligned}
P_B(6 \leq X \leq 10) &\approx P_N(5.5 \leq X \leq 10.5) \\
&= P_N\left(\frac{5.5-8}{2} \leq \frac{X-8}{2} \leq \frac{10.5-8}{2}\right) \\
&= P_N(-1.25 \leq Z \leq 1.25) \\
&= P(Z \leq 1.25) - P(Z \leq -1.25) \\
&= 0.8944 - 0.1056 \\
&= 0.7888
\end{aligned}$$

That is, the normal approximation for $P_B(6 \leq X \leq 10)$ is 0.7888. The exact value of $P_B(6 \leq X \leq 10)$ is 0.7898.

The reason for using 5.5 and 10.5 rather than 6 and 10 in P_N (usually called *half-integer corrections for continuity*) is evident when we look at Figure 5.7.1. If we graph the probabilities given by the binomial probability function $b(x)$ by drawing rectangles having bases equal to 1 and *centered* at $x = 0, 1, \ldots, 16$ and heights given by $b(0), b(1), \ldots, b(16)$, the area under the resulting probability *histogram* is 1 since each rectangle is of area $b(x) \times 1$ and

$$\sum_{x=0}^{16} b(x) = 1$$

The probability histogram is often called the *probability bar chart.*

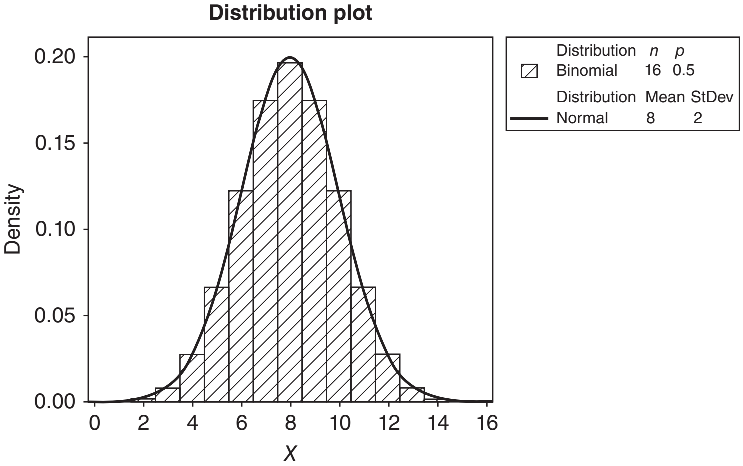


Figure 5.7.1 Graph of the binomial probability histogram in Example 5.7.1 and of the approximating normal distribution $N(8, 4)$.

When computing $P_B(6 \leq X \leq 10) = \sum_{x=6}^{10} b(x)$, we are summing areas of rectangles, the first of which has a base with left-hand endpoint 5.5 and the last of which has a base with right-hand endpoint 10.5. Now, if we approximate $P_B(6 \leq X \leq 10)$ by $P_N(6 \leq X \leq 10)$, that is, if we do not make the half-integer correction, we are, in effect, omitting about half of the first and half of the last rectangle of the probability bar charts from consideration, thus underestimating the required probability.

We can also approximate individual binomial probabilities. For instance,

$$\begin{aligned} P_B(X = 8) &\approx P_N(7.5 \leq X \leq 8.5) \\ &= P_N(-0.5/2 \leq Z \leq 0.5/2) \\ &= P_N(-0.25 \leq Z \leq 0.25) \\ &= 0.5987 - 0.4013 \\ &= 0.1974 \end{aligned}$$

Using binomial probability tables, we can see that the exact value of $P_B(X = 8)$ is 0.1964. Thus, we see from Example 5.7.1 that the procedure to approximate binomial probabilities by the normal distribution involves setting the mean and variance of the binomial distribution equal to the mean and variance of the approximating normal distribution and then making the needed half-integer corrections for continuity.

Example 5.7.2 (Monitoring defects using normal approximation) *A machine produces items in a certain plant using a mass-production process. Its record is that 5% of the items produced are defective. (a) If a sample of 1000 items is chosen at random, what is the probability that no more than 40 defectives occur in the sample? (b) that between 40 and 60, inclusive, of the items in the sample are defective?*

Note here $\mu = np = 1000 \times (1/20) = 50$, $\sigma^2 = np(1-p) = 95/2 = 47.5$, *and* $\sigma = 6.892$.

Solution: Let X= the number of defectives in the sample of 1000 items. Since $np = 50 > 5$ and $n(1-p) = 950 > 5$, using the normal approximation, we find:

(a)

$$\begin{aligned} P_B(X \leq 40) &\approx P_N(X \leq 40.5) \\ &= P_N\left(Z \leq \frac{40.5 - 50}{6.892}\right) \\ &= P_N(Z \leq -1.38) \\ &= 0.084 \end{aligned}$$

(b)

$$\begin{aligned} P_B(40 \leq X \leq 60) &\approx P_N(39.5 \leq X \leq 60.5) \\ &= P_N\left(\frac{-10.5}{6.892} \leq Z \leq \frac{10.5}{6.892}\right) \\ &= P_N(-1.52 \leq Z \leq 1.52) \\ &= 0.871 \end{aligned}$$

5.7.2 Approximation of the Poisson Distribution by the Normal Distribution

If X is a Poisson random variable with mean and variance λ, we then have the result given below.

$$P(a \le X \le b) \approx P\left(\frac{(a-0.5)-\lambda}{\sqrt{\lambda}} \le Z \le \frac{(b+0.5)-\lambda}{\sqrt{\lambda}}\right) \tag{5.7.2}$$

where a and b are nonnegative integers, Z is a standard normal random variable, and ±0.5 is a continuity correction factor. The continuity correction factor is applied in the same manner as for the binomial distribution.

5.8 A TEST OF NORMALITY

Normal probability graph paper has the property that the graph of the c.d.f. of a normal distribution, when plotted on this special paper, is a straight line. That is, by an appropriate stretch of the scale of ordinates for low and high percentages, the graph of the c.d.f. of an $N(\mu, \sigma^2)$ random variable, as shown in Figure 5.5.5, is transformed into a straight line as shown in Figure 5.8.1.

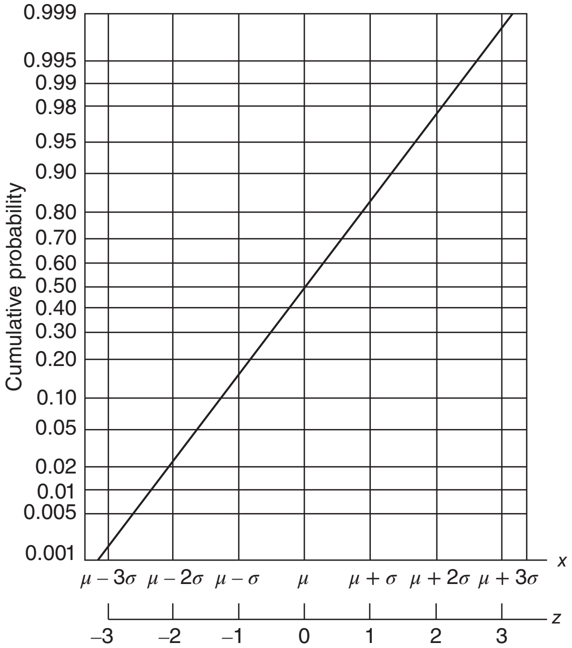


Figure 5.8.1 Graph of the cumulative normal distribution function on probability graph paper.

Normal probability paper provides a rough check on whether a sample can reasonably be regarded as having come from a normal population. We illustrate the use of normal probability paper with the following example.

Example 5.8.1 (A graphical test of normality) *The data below are the systolic blood pressures of 16 randomly selected breast cancer patients in a large metropolitan hospital. Using normal probability paper, we want to verify if these data can reasonably be regarded as having come from a normal population.*

134,138, 133,125, 128,123, 130,134, 114,136, 124,146, 147,119, 133,135

Solution: We plot the normal probability paper using the following steps:

Step 1. Arrange the data in ascending order, so that we have 114,119, 123,125, 125,128, 130,133, 133,134, 134,135, 136,138, 146,147.

Step 2. Compute $p_i = (i - 1/2)/n$, where i takes values $1, \ldots, n$; n is the total number of observations, and p_i is the ith cumulative probability corresponding to the ith-ordered observation (see Step 1 above).

Step 3. Plot the ith ordered observation on the horizontal axis and the corresponding p_i on the vertical axis.

Note that if the data comes from a normal population, then all the points plotted on the normal probability paper should fall on, or be close to, a straight line. The normal probability plot for the data in this example is shown in Figure 5.8.2, which clearly indicates that the plotted data points fall either on or close to a straight line. Thus, we conclude that these data come from a normal population. Furthermore, if all the points plotted on the normal probability paper fall on or very close to the straight line, then we use the 50th percentile of the sample to estimate the mean μ and the difference between the 84th

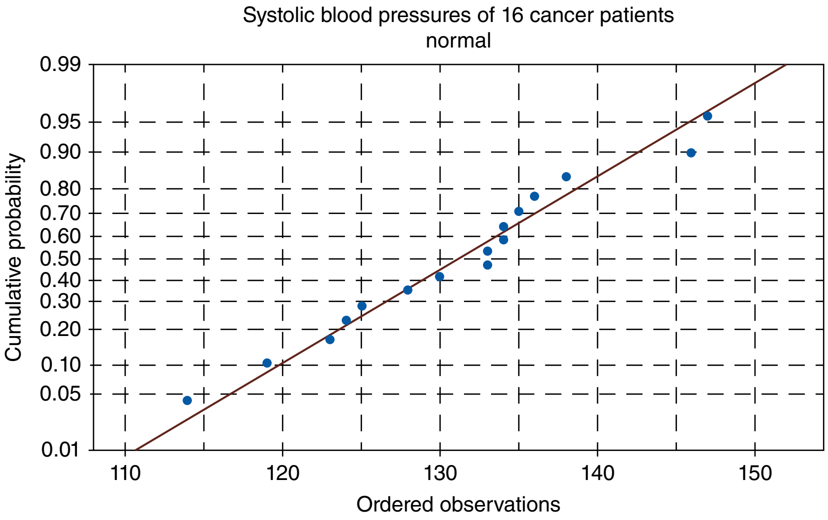


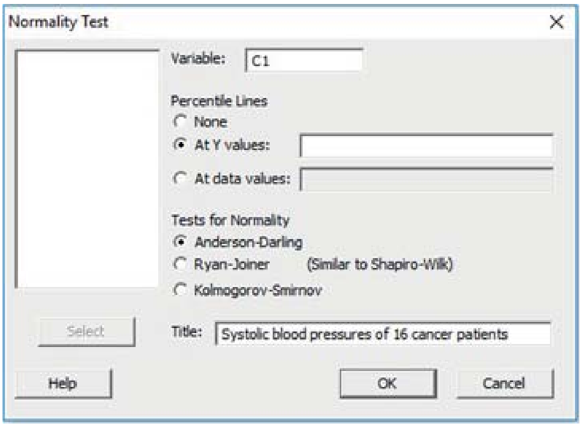
Figure 5.8.2 Normal probability plot for the data in Example 5.8.1.

and 50th percentiles to estimate the standard deviation σ. For example, from Figure 5.8.2, we get an estimate of the mean μ and standard deviation σ as 131 and $140 - 131 = 9$, respectively.

We may also test if the data in Example 5.8.1 can reasonably be regarded has having come from a normal population using MINITAB and R. In order to do so, we proceed as follows:

MINITAB

1. Enter the 16 values in column C1.
2. From the Menu bar, select **Stat** > **Basic Statistics** > **Normality Test**.
3. The following dialog box appears on the screen.



Normality Test

Variable: C1

Percentile Lines

○ None

● At Y values:

○ At data values:

Tests for Normality

● Anderson-Darling

○ Ryan-Joiner (Similar to Shapiro-Wilk)

○ Kolmogorov-Smirnov

Title: Systolic blood pressures of 16 cancer patients

4. Enter C1 in the box next to **variable** and select the circle next to **At Y Values**.
5. Under **Test for Normality** check the circle next to one of the three possible tests, say **Anderson-Darling**.
6. If you like, put the title of the problem in the box next to **Title** and Click **OK**. Then in the session window, the Normal Probability Graph will appear as shown in Figure 5.8.3.

The small box in the right hand corner in Figure 5.8.3 provides various statistics. The last statistic gives a so-called p-value (to be studied in Chapter 9). In general, if the p-value is > 0.05, we have sufficient evidence to assume that the data come from a normal population. Thus, we can say that the data in Example 5.8.1 come from a normal population.

USING R

In R, the built in function ‘qqnorm()’ can be used to plot the normal quantile-quantile plot (q-q plot). The x-axis of this graph shows the observed sample quantiles, while the y-axis shows the theoretical standard normal quantiles. Exclusion of the option ‘datax=TRUE’ would swap the x and y axes. The additional option ‘main’ can be used to get the title if needed, and the options ‘pch’, ‘cex’, and ‘col’ can be used to get shaded, double-sized colored points. An additional straight line ‘qqline()’ can be added to visualize the deviations of the observed data from the normal distribution. The options ‘col = 2’, ‘lwd = 2’, and ‘lty =2’ are used to get a red-colored, double-sized, dashed line. However, this line is slightly different from Minitab as Minitab generates the line based on the mean and standard deviation of a normal distribution fitted to the data, but R draws a line between the

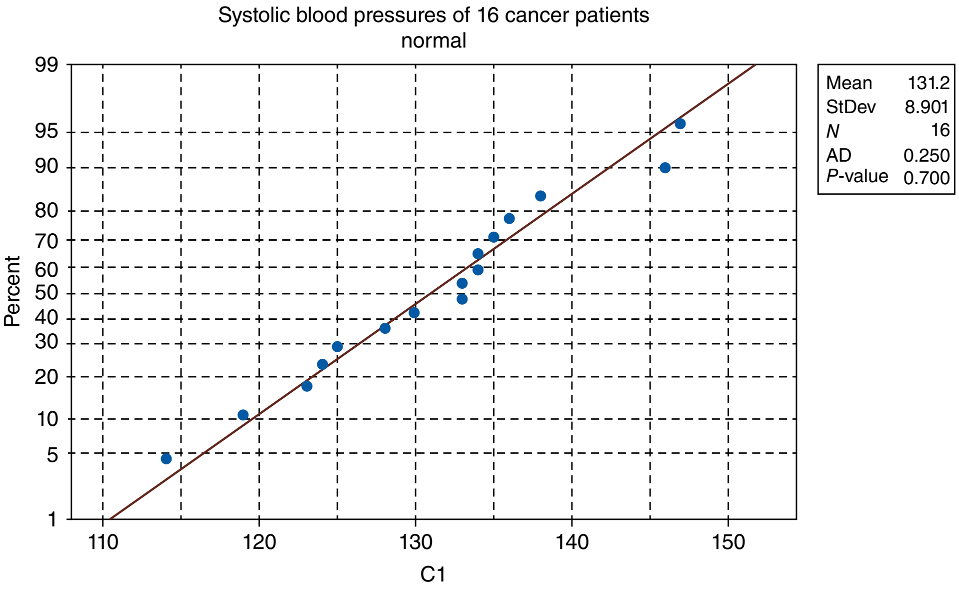


Figure 5.8.3 MINITAB normal probability plot for the data in Example 5.8.1.

first and third quantiles for the sample and theoretical quantiles. The normal probability plot for the data in Example 5.8.1 is shown in Figure 5.8.4 and can be obtained by typing the following commands in the R Console window.

```
x = c(134,138,133,125,128,123,130,134,114,136,124,146,147,119,133,135)
qqnorm(x, datax=TRUE, main=“Systolic blood pressures of 16 cancer patients”,
pch=20, col=4, cex=1.5)
qqline(x, col = 2, lwd=2, lty=2, datax=TRUE)
```

Again, from Figure 5.8.4, we can conclude that the data from Example 5.8.1 comes from a normal population because the plotted points fall almost on a straight line.

PRACTICE PROBLEMS FOR SECTIONS 5.7 AND 5.8

1. The following data give the time to the nearest minute needed for a computer to execute certain simulation problems:

30	28	35	28	36	29	35	24
26	25	21	25	22	22	26	28

Use normal probability paper (or one of the statistical packages) to verify whether or not this data set comes from a normal population. Find the sample mean and sample standard deviation.

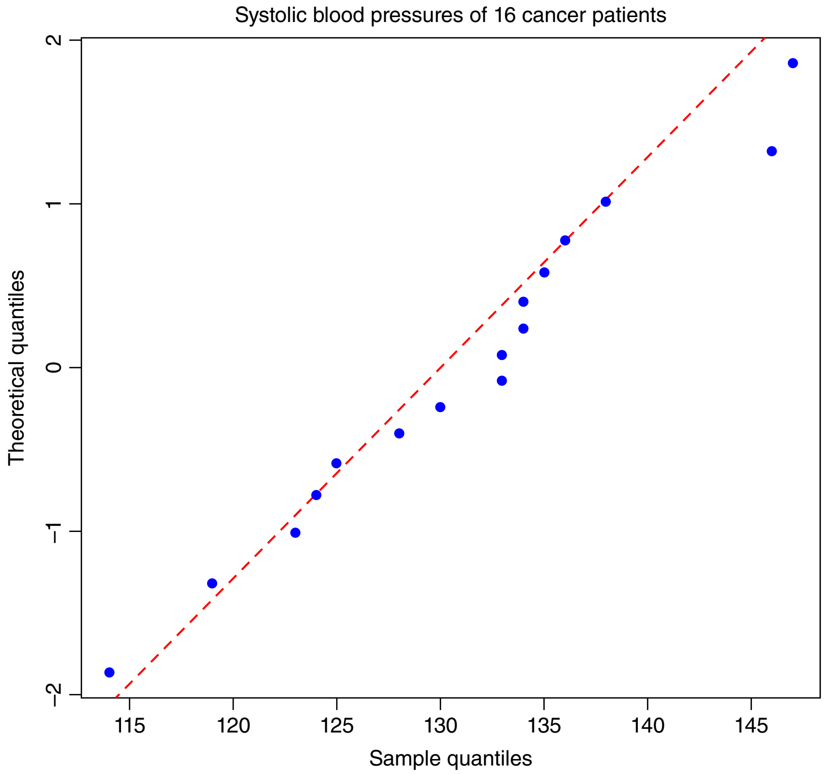


Figure 5.8.4 R normal probability plot for the data in Example 5.8.1.

2. The following data give the amount of effluent water discharged per second from a sewage treatment facility measured over 20 days:

61	61	76	67	76	62	67	67	70	70
69	66	76	73	69	69	70	73	65	65

Use normal probability paper (or one of the statistical packages) to verify whether or not these data come from a normal population. Find the sample mean and sample standard deviation.

3. Use normal probability paper (or one of the statistical packages) to verify whether or not the following data come from a normal population:

11	14	6	12	11	10	8	8	11	8

4. Referring to Problem 3, if you conclude that the data come from a normal population, then use normal probability paper to estimate the mean and the standard deviation of the population.

5. Let X be a binomial random variable with parameters $n = 20$ and $p = 0.4$.
   (a) Use the binomial tables (Table A.2) to determine $P(4 \leq X \leq 12)$.
   (b) Find the mean and the standard deviation of the binomial distribution, and then use the appropriate normal approximation to find $P(4 \leq X \leq 12)$. Compare the approximate probability with the actual probability you determined in (a).
6. Find the following probabilities using the normal approximation to the binomial distribution with parameters $n = 100$, $p = 0.8$: (a) $P(73 \leq X \leq 87)$, (b) $P(73 < X < 88)$, (c) $P(X \geq 81)$, (d) $P(X < 77)$.
7. A semiconductor manufacturer recently found that 3% of the chips produced at its new plant are defective. Assume that the chips are independently defective or nondefective. Use a normal approximation to determine the probability that a box of 500 chips contains: (a) at least 10 defective chips (b) between 15 and 20 (inclusive) defective chips.
8. A pathology lab does not deliver 10% of the test results in a timely manner. Suppose that in a given week, it delivered 400 test results to a certain hospital. Use the normal approximation to the binomial to find the following probabilities, assuming that test results delivered are independent of one another.
   (a) At least 30 test results were not delivered in a timely manner.
   (b) Between 25 and 35 (inclusive) test results were not delivered in a timely manner.
   (c) At most 40 test results were not delivered in a timely manner.
9. An European airline company flies jets that can hold 200 passengers. The company knows from past experience that on the average, 8% of the booked passengers do not show up on time to take their flight. If the company booked 212 passengers for a particular flight, then what is the probability that the plane will have some empty seats? Assume that the passengers for that flight show up independently.
10. A random variable X is distributed as binomial with parameters $n = 20, p = 0.4$. Compute the exact and the normal approximation to $P(10 \leq X \leq 16)$ and compare the two probabilities.

5.9 PROBABILITY MODELS COMMONLY USED IN RELIABILITY THEORY

An important industrial application of statistics concerns the reliability and life testing of manufactured devices or systems. Reliability is defined as the probability that a device or system performs properly, for the period of time intended, under acceptable operating conditions. A life test is done by exposing a device to an environment typical of the acceptable operating conditions and noting the time to failure. Accelerated life tests occur when extreme environment stresses are applied. A subject closely allied to reliability is that of the actuarial sciences wherein histories of the times to death of individuals are used to determine insurance strategies.

Here, we study the four distributions that are used quite often in reliability theory, namely the *lognormal distribution, exponential distribution, gamma distribution*, and the *Weibull distribution*. We will discuss reliability theory later, in Chapter 10.

5.9.1 The Lognormal Distribution

Definition 5.9.1 A random variable X is said to be distributed as lognormal with parameters μ and σ^2 if the random variable $Y = \ln X$ is distributed as $N(\mu, \sigma^2)$. The p.d.f. of the random variable X is given by

$$f(x) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, & x > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{5.9.1}$$

Alternatively, we say that the random variable $X = e^Y$ is distributed as lognormal with parameters μ and σ^2 if Y is $N(\mu, \sigma^2)$. The graph in Figure 5.9.1 gives the shapes of lognormal density function for various values of σ and $\mu = 0$. The reader should note that μ and σ^2 are the mean and variance of the random variable Y and *not* of X.

We may also describe a lognormal random variable as follows: if a random variable Y is distributed as $N(\mu, \sigma^2)$, then the random variable $X = e^Y$ is distributed as lognormal so that

$$E(X) = E(e^Y) \quad \text{and} \quad Var(X) = Var(e^Y),$$

where Y is a $N(\mu, \sigma^2)$ random variable.

Now using (5.2.1) and (5.2.3), we state the *mean* and *variance* of the lognormal distribution with parameters μ and σ^2 in the results (5.9.2) and (5.9.3) stated below.

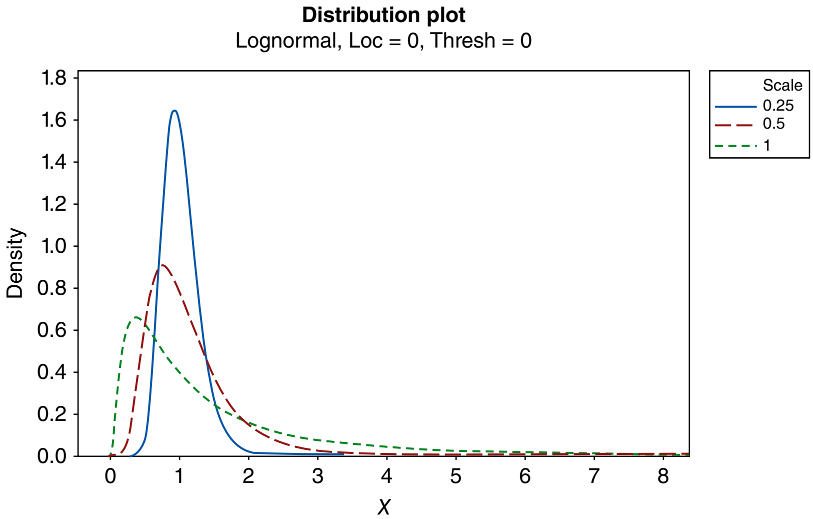


Figure 5.9.1 MINITAB graphs of lognormal density function for $\sigma = 0.25, 0.5$, and 1.0, with $\mu = 0$.

Mean and *variance* of the lognormal distribution with parameters μ and σ^2:

$$E(X) = \exp(\mu + \sigma^2/2) \tag{5.9.2}$$

$$Var(X) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) \tag{5.9.3}$$

These results can be obtained using the m.g.f. of the normal distribution, that is,

$$E(e^{tY}) = \exp(\mu t + (1/2)\sigma^2 t^2),$$

where Y is distributed as $N(\mu, \sigma^2)$, and that $E(X) = E(e^Y)$, and $Var(X) = Var(e^Y) = E(e^Y)^2 - (E(e^Y))^2 = E(e^{2Y}) - (E(e^Y))^2$.

Example 5.9.1 (Mean and variance of lognormal random variable) *The random variable* X *denotes the lifetime of an active ingredient in a drug that is modeled by the lognormal distribution with* $\mu = 1.20$ *yr and* $\sigma^2 = 0.10\ (\text{yr})^2$. *Find the mean and the standard deviation of the lifetime of the active ingredient.*

Solution:

$$\begin{aligned} Mean = E(X) &= \exp(\mu + \sigma^2/2) \\ &= e^{1.20+0.10/2} = e^{1.25} \\ &= 3.49 \text{ yr} \\ Var(X) &= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) \\ &= e^{2.40+0.20} - e^{2.40+0.10} = e^{2.60} - e^{2.50} \\ &= 1.28\ (\text{yr})^2 \end{aligned}$$

Therefore, the standard deviation is 1.13 year.

The probabilities for the lognormal distribution can be found using tables of the standard normal distribution, as illustrated by the following example.

Example 5.9.2 (Lognormal probabilities) *Refer to Example 5.9.1. Find the following probabilities: (a)* $P(X \geq 4)$, *(b)* $P(X \leq 2)$, *(c)* $P(3 \leq X \leq 5)$.

Solution: Since $\ln X$ is distributed as $N(1.20, 0.10)$, to find the probabilities above, we proceed as follows:

(a)

$$\begin{aligned} P(X \geq 4) &= P(\ln X \geq \ln 4) \\ &= P(\ln X \geq 1.386) \\ &= P\left(\frac{\ln X - 1.20}{0.316} \geq \frac{1.386 - 1.20}{0.316}\right) \\ &= P(Z \geq 0.59) \approx 0.2776. \end{aligned}$$

(b)

$$\begin{aligned}P(X \leq 2) &= P(\ln X \leq \ln 2)\\ &= P(\ln X \leq 0.693)\\ &= P\left(\frac{\ln X - 1.20}{0.316} \leq \frac{0.693 - 1.20}{0.316}\right)\\ &= P(Z \leq -1.60) \approx 0.0548.\end{aligned}$$

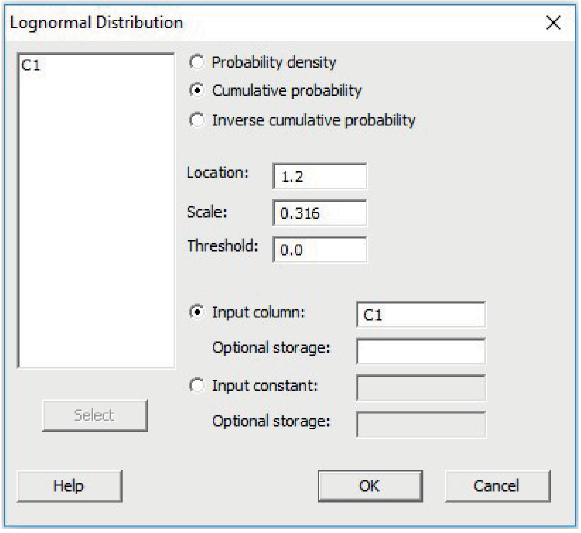
(c)

$$\begin{aligned}P(3 \leq X \leq 5) &= P(\ln 3 \leq \ln X \leq \ln 5)\\ &= P(1.0986 \leq \ln X \leq 1.6094)\\ &= P\left(\frac{1.0986 - 1.20}{0.316} \leq \frac{\ln X - 1.20}{0.316} \leq \frac{1.6094 - 1.20}{0.316}\right)\\ &= P(-0.32 \leq Z \leq 1.30) \approx 0.5287.\end{aligned}$$

The reader can also determine these probabilities using MINITAB or R. For example, to find the probability in Part (c) above, by proceeding as follows:

MINITAB

1. Enter the values 3 and 5 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **Lognormal**.
3. In the dialog box, click the circle next to **Cumulative probability**.



Lognormal Distribution

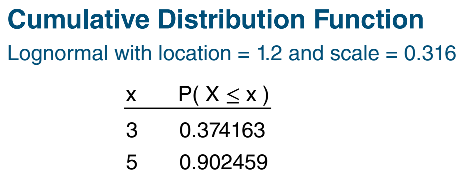
C1

○ Probability density
● Cumulative probability
○ Inverse cumulative probability

Location: 1.2
Scale: 0.316
Threshold: 0.0

● Input column: C1
Optional storage:
○ Input constant:
Optional storage:

4. Enter 1.20 (the value of μ) in the box next to **Location**, 0.316 (the value of σ) next to **Scale**, and zero next to **Threshold**.
5. Click the circle next to **Input column** and type C1 in the box next to it.
6. Click **OK**.
7. In the session window, text appears as shown in the following box. Thus, $P(3.0 \leq X \leq 5.0) = P(X \leq 5.0) - P(X \leq 3.0) = 0.902459 - 0.374163 = 0.5283$.



```
Cumulative Distribution Function
Lognormal with location = 1.2 and scale = 0.316

x    P( X ≤ x )
3    0.374163
5    0.902459
```

USING R

R has a built in cumulative lognormal distribution function 'plnorm(q, meanlog, sdlog)', where q is the quantile and meanlog and sdlog are the mean and the standard deviation of the lognormal distribution, respectively. So, referring to Example (5.9.1), we are asked to find $P(3.0 \leq X \leq 5.0)$, which can be computed by typing the following in the R Console window.

```
plnorm(5, 1.2, 0.316) - plnorm(3, 1.2, 0.316)

#R output
[1] 0.5282957
```

We find that $P(3.0 \leq X \leq 5.0) \approx 0.5283$, the same value found using MINITAB. We now state another important result related to lognormal random variables.

Let X_i, $i = 1, 2, \ldots, n$, be n independent lognormal random variables with parameters μ_i and σ_i^2; then, the random variable $Y = \prod_{i=1}^{n} X_i$ is also lognormal with parameters $\mu = \mu_1 + \cdots + \mu_n$ and $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$. This result follows immediately by noting that $\ln Y = \sum_{i=1}^{n} \ln X_i$ and then using Definition 5.9.1 and the result of Theorem 5.6.1.

PRACTICE PROBLEMS FOR SECTION 5.9.1

1. The lifetime, in hours, of a head gasket of a cylinder in a car engine is distributed as lognormal with mean 6000 and standard deviation 5000 hours. Find the probability that the lifetime of the head gasket is more than 8000 hours.
2. The size of a computer chip is distributed as lognormal with parameters μ and σ The following data give the sizes of eight randomly selected chips:

4.18	2.83	3.76	4.79	3.59	2.98	4.16	2.12

   Estimate μ and σ.
3. Suppose that a random variable X is distributed as lognormal with parameters $\mu = 2$ and $\sigma = 0.5$. Find the mean and the variance of the random variable X.
4. Suppose that a random variable X is normally distributed with $\mu = 4$ and $\sigma^2 = 4$. If the lifetime Y (hours), of an electronic component can be modeled with the

distribution of the random variable $Y = e^X$, then find the following: (a) the mean and the standard deviation of Y, (b) $P(Y > 4000)$, (c) $P(Y \leq 5000)$.

5. Suppose that a random variable Y has a lognormal distribution with parameters $\mu = 4$ and $\sigma^2 = 4$. Determine the following: (a) the mean and standard deviation of Y, (b) $P(Y > 250)$, (c) $P(100 < Y < 200)$.
6. Suppose that the lifetime X, in hours, of a surgical instrument can be modeled with lognormal distribution having parameters $\mu = 8$ and $\sigma^2 = 2$. Determine the following: (a) $P(X \geq 25{,}000)$, (b) $P(X > 24{,}000)$, (c) $P(22{,}000 < X < 28{,}000)$.

5.9.2 The Exponential Distribution

In Chapter 4, we studied the Poisson probability distribution, which describes the phenomenon of random events that occur in a Poisson process. The events in the Poisson process occur randomly. This suggests, for example, that the time between the occurrences of any two consecutive events, say T, is a random variable. The random variable T is then distributed as exponential. As another example, the distance between two defects in a telephone or electric wire is distributed as an exponential when these defects follow the Poisson process. The exponential distribution has a wide range of applications in any process that does not take the aging/anti-aging factor into account. For example, if a machine is always as good as new, then to study its reliability, we use the exponential distribution.

Definition 5.9.2 A random variable X is said to be distributed by the *exponential distribution with parameter* λ, if its p.d.f. is defined as follows:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for} \quad x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.9.4}$$

In Definition 5.9.2, $\lambda > 0$ is the only parameter of the distribution and $e \approx 2.71828$. It is important to note that λ is the number of events occurring per unit time, per unit length, per unit area, or per unit volume in a Poisson process. The shape of the density function of an exponential distribution changes as the value of λ changes. Figure 5.9.2 shows density functions of exponential distributions for some selected values of λ.

Now from (5.2.7) and (5.2.8), the mean and standard deviation of the exponential distribution with p.d.f. given by (5.9.4) are

$$\mu = E(X) = 1/\lambda \quad \text{and} \quad \sigma = \sqrt{Var(X)} = 1/\lambda \tag{5.9.5}$$

The shape of the density function of a random variable X distributed as exponential distribution changes as the value of the parameter λ changes (see Figure 5.9.2).

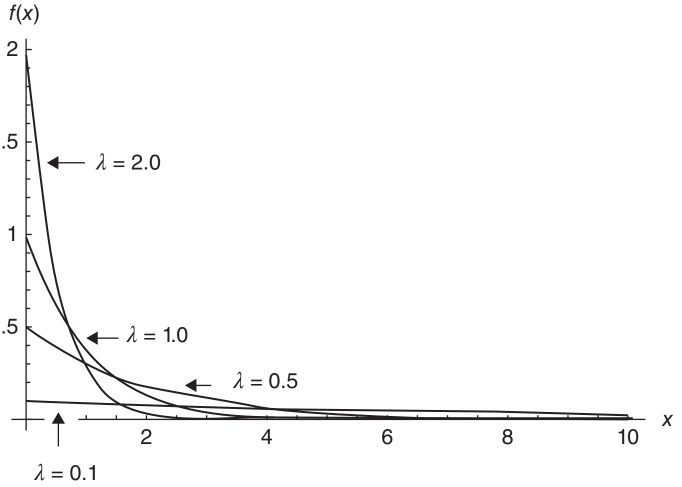


Figure 5.9.2 Graphs of exponential density function for $\lambda = 0.1, 0.5, 1.0$, and 2.0.

Distribution Function $F(x)$ of the Exponential Distribution

Consulting (5.9.4), an exponential random variable with parameter λ has its c.d.f., for $x > 0$, given by

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_0^x \lambda e^{-\lambda t} dt \\ &= 1 - e^{-\lambda x} \end{aligned} \tag{5.9.6}$$

It follows that

$$\begin{aligned} P(X > x) &= 1 - P(X \leq x) \\ &= 1 - F(x) \\ &= 1 - (1 - e^{-\lambda x}) = e^{-\lambda x} \end{aligned} \tag{5.9.7}$$

Equation (5.9.7) leads us to an important property known as the *memoryless* property of the exponential distribution. As an illustration of this property, we consider the following example.

Example 5.9.3 (Memoryless property of the exponential distribution) *Let the breakdowns of a machine follow the Poisson process so that the random variable* X *denoting the number of breakdowns per unit of time is distributed as a Poisson distribution. Then, the time between any two consecutive failures is also a random variable* T *(say) that is distributed as an exponential with parameter* λ*. Assuming* $\lambda = 0.1$*, determine the following probabilities:*

(a) $P(T > t)$*, that is, the probability that the machine will function for at least time* t *before it breaks down again.*

(b) $P(T > t + t_1|T > t_1)$. Note that the condition $T > t_1$ means that it is known the machine has already operated for time t_1 after a breakdown, and then we wish to find the probability it will function for at least t *more units of time starting from time t_1. Hence, we wish to find the probability that the machine will function for a total time of at least* $\mathrm{t} + t_1$ *before its next breakdown given that it has already functioned for time t_1 since the previous breakdown. Here, t_1 and t are positive numbers.*

Solution: (a) Given that $\lambda = 0.1$, we want to find the probability $P(T > t)$. From (5.9.7), this probability is given by

$$P(T > t) = e^{-\lambda t} = e^{-(0.1)t} = e^{-t/10}$$

(b) In this case, we are interested in finding the conditional probability $P(T > t + t_1|T > t_1)$. Using the definition of conditional probability, we wish to find

$$P(T > t + t_1|T > t_1) = \frac{P[(T > t + t_1) \cap (T > t_1)]}{P(T > t_1)}$$

But the event $[(T > t + t_1) \cap (T > t_1)]$ means that $T > t + t_1$ and $T > t_1$. This implies that $(T > t + t_1)$, since $t > 0$ and $T > t_1 > 0$. Thus, we have

$$P(T > t + t_1|T > t_1) = \frac{P(T > t + t_1)}{P(T > t_1)}$$

Now using the result in (5.9.7), we have

$$\begin{aligned} P(T > t + t_1|T > t_1) &= \frac{e^{-\lambda(t+t_1)}}{e^{-\lambda t_1}} \\ &= e^{-\lambda t} \\ &= e^{-t/10} \end{aligned}$$

since $\lambda = 0.1$. Therefore, the probability $P(T > t + t_1|T > t_1)$ is the same as the probability $P(T > t)$. This result means that under the exponential model, the probability $P(T > t)$ remains the same no matter from what point we measure the time t. In other words, it does not remember when the machine has had its last breakdown. For this reason, the exponential distribution is known to have a *memoryless property.*

From the previous discussion, we can see that it does not matter when we start observing the system since it does not take into account an aging factor. That is, whether the machine is brand new or 20 years old, we have the same result (or exponential model) as long as we model the system using the same Poisson process. In practice, however, this conclusion is not completely valid. For example, if we are investigating how toll booths function during rush hours and nonrush hours, and we model it with the same Poisson process, then the results may not be valid because of the two different time periods in the traffic flow. It would make more sense that when there is clear distinction between the two scenarios, we should model them by two different processes.

It is easy to see that the m.g.f. of the exponential random variable X is as given in (5.9.8) below.

Moment-generating function of the exponential random variable X:

$$M_X(t) = \lambda \int_0^{\infty} e^{-x(\lambda-t)}dx = (1 - t/\lambda)^{-1} \quad \text{for} \quad \lambda > t \tag{5.9.8}$$

Now by taking the first and second derivative of $M_X(t)$ in Equation (5.9.8) and then substituting $t = 0$, it is easy to prove that the mean and the variance of the exponential distribution (5.9.4) is given below.

Mean and variance of the exponential distribution:

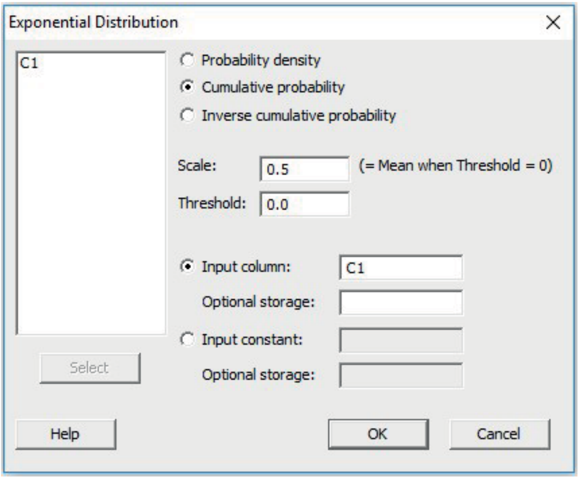
$$\mu = E(X) = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = Var(X) = \frac{1}{\lambda^2} \tag{5.9.9}$$

Example 5.9.4 (Using MINITAB and R for exponential probabilities) *Let* X *be a random variable distributed as an exponential distribution with* $\lambda = 2$ *or mean value equal to 0.5, which is the scale parameter. Determine the following probabilities using MINITAB and R: (a)* $P(X \geq 3)$*, (b)* $P(X \leq 5.0)$*, (c)* $P(3.0 \leq X \leq 5.0)$*.*

Solution: In order to determine the probabilities $P(X \geq 3.0)$, $P(X \leq 5.0)$, and $P(3.0 \leq X \leq 5.0)$, we have to first find the probabilities $P(X \leq 3.0)$ and $P(X \leq 5.0)$. Then, $P(X \geq 3.0) = 1 - P(X \leq 3.0)$, $P(3.0 \leq X \leq 5.0) = P(X \leq 5.0) - (X \leq 3.0)$. To find the probabilities $P(X \leq 3.0)$ and $P(X \leq 5.0)$, we proceed as follows:

MINITAB

1. Enter the values 3 and 5 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **Exponential**.
3. In the following dialog box, click the circle next to **Cumulative probability**.



Exponential Distribution

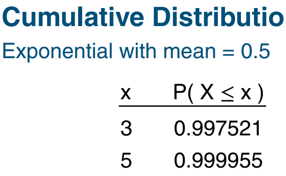
C1

○ Probability density
● Cumulative probability
○ Inverse cumulative probability

Scale: 0.5 (= Mean when Threshold = 0)
Threshold: 0.0

● Input column: C1
Optional storage:
○ Input constant:
Optional storage:

4. Enter 0.5 (the value of the mean) in the box next to **Scale**, since the threshold parameter is zero.
5. Click the circle next to **Input column** and type C1 in the box next to it and Click **OK**.
6. In the Session window, text will appear as shown below, in the following box. Thus, $P(X \geq 3.0) = 1 - P(X \leq 3.0) = 1 - 0.997521 = 0.002479; P(X \leq 5.0) = 0.999955$, so that $P(3.0 \leq X \leq 5.0) = P(X \leq 5.0) - P(X \leq 3.0) = 0.999955 - 0.997521 = 0.002434$



```
Cumulative Distribution Function
Exponential with mean = 0.5

x    P( X ≤ x )
3    0.997521
5    0.999955
```

USING R

R has a built in cumulative exponential distribution function 'pexp(q, rate)', where q is the quantile and rate is the rate parameter which is equal to 1/mean of the exponential distribution. So, referring to Example 5.9.4, we are asked to find $P(3.0 \leq X \leq 5.0)$, which can be computed by typing the following in the R Console window.

```
pexp(5, 2) - pexp(3, 2) #Note that: rate = 1/0.5 = 2

#R output
[1] 0.002433352
```

PRACTICE PROBLEMS FOR SECTION 5.9.2

1. The waiting time T at a bank teller's window between two successive customers is distributed as exponential with a mean of four minutes. Find the following probabilities:
   (a) $P(T \geq 5$, (b) $P(3 \leq T \leq 6)$, (c) $P(T \leq 4)$, and (d) $P(T < 5)$.
2. Suppose that a random variable X is distributed as exponential distribution with parameter λ. Show that $P(X \leq 10) = P(X \leq 17 | X \geq 7)$. This property of exponential distribution is known as the "memoryless" property.
3. Suppose that the lapse of time between two successive accidents in a paper mill is exponentially distributed with a mean of 15 days. Find the probability that the time between two successive accidents at that mill is more than 20days.
4. Let a random variable X be distributed exponentially with mean equal to 10. Find the m.g.f. of the random variable $Y = 4 + 3X$. Then, use this m.g.f. to find the mean and variance of the random variable Y.
5. Suppose that a computer lab in a small liberal arts college has 20 similar computers. Let the random variable T denote the time, in years, to failure of this type of computer. If the random variable T follows the exponential distribution with mean equal to three years, then find the probability that 15 of the 20 computers are still functioning after five years.

6. The lifetime X, in years, of car batteries has an exponential distribution with a mean life of five years. If you buy a new car and plan to keep it for six years, then find the probability that you will change the battery during your ownership.
7. The time T, in minutes, between the arrival of two successive patients in an emergency room can be modeled as an exponential distribution with mean 20 minutes. Determine the following probabilities:
   (a) $P(T > 30)$, (b) $P(12 < T < 18)$, (c) $P(T < 25)$.
8. Suppose that the time between arrivals of two buses at a city bus station is exponentially distributed with a mean of five minutes. Determine the following probabilities:
   (a) More than two buses arrive during an interval of 5 minutes.
   (b) No bus arrives during an interval of 10 minutes.
   (c) No more than two buses arrive during an interval of 15 minutes.

5.9.3 The Gamma Distribution

Another probability distribution found useful in reliability theory is the *gamma distribution*. The gamma distribution is frequently used as a probability model for waiting times. For example, in life testing, the waiting time until death is a random variable that is usually modeled with a gamma distribution.

Definition 5.9.3 The random variable T is said to be a *gamma variable of order* γ with parameter λ if its p.d.f. is given by

$$f(t|\gamma,\lambda) = \begin{cases} \dfrac{\lambda^\gamma}{\Gamma(\gamma)} t^{\gamma-1} e^{-\lambda t}, & \text{if } t \geq 0 \\ 0, & \text{if } t < 0 \end{cases} \tag{5.9.10}$$

for $\gamma > 0, \lambda > 0$.

Note that γ is sometimes called the *shape parameter*, and $1/\lambda$ is called the *scale parameter*. The shape of the gamma p.d.f. (5.9.10) changes as the parameters λ and γ change. The p.d.f. of the gamma distribution for various values of the shape parameter γ and the scale parameter $\lambda = 1$ is shown in Figure 5.9.3.

Further note that $\Gamma(\gamma)$ is defined as

$$\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} e^{-x} dx, \quad \gamma > 0 \tag{5.9.11}$$

When the random variable T must originate at some *threshold parameter* value $\tau_c > 0$, the gamma distribution is written as

$$f(t|\gamma,\lambda,\tau_c) = \begin{cases} \dfrac{\lambda^\gamma}{\Gamma(\gamma)} (t-\tau_c)^{\gamma-1} e^{-\lambda(t-\tau_c)}, & \text{if } \ t \geq \tau_c \\ 0, & \text{if } \ t < \tau_c \end{cases} \tag{5.9.12}$$

We remark that it can easily be shown that

$$\Gamma(\gamma) = (\gamma - 1)\Gamma(\gamma - 1) \tag{5.9.13}$$

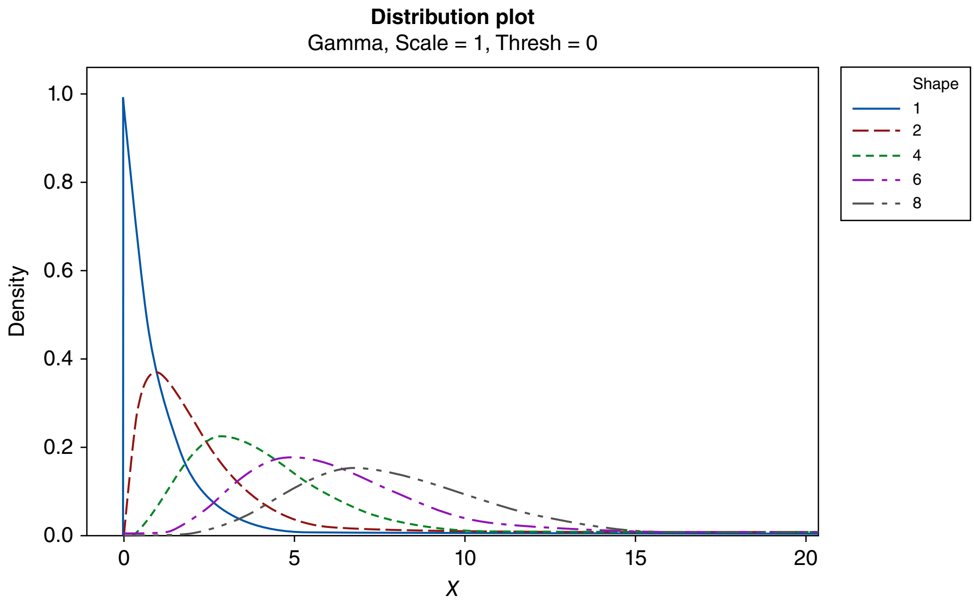


Figure 5.9.3 Probability density function of the gamma distribution for various values of the shape parameter γ and the scale parameter $\lambda = 1$.

so that if γ is a positive integer

$$\Gamma(\gamma) = (\gamma - 1)! \tag{5.9.14}$$

and

$$\Gamma(1) = 0! = 1 \tag{5.9.15}$$

The m.g.f. of the gamma random variable X is given below.

Moment-generating function of the gamma distribution:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\gamma} \tag{5.9.16}$$

Now, by taking the first and second derivative of $M_X(t)$ in Equation (5.9.16) and then substituting $t = 0$, it is easy to prove that the mean and the variance of the gamma distribution (5.9.10) are as given in (5.9.17) below.

Mean and variance of the gamma distribution:

$$\mu = E(X) = \gamma/\lambda \quad \text{and} \quad \sigma^2 = Var(X) = \gamma/\lambda^2 \tag{5.9.17}$$

We note that the gamma distribution (5.9.10) becomes the exponential distribution when $\gamma = 1$.

Example 5.9.5 (Gamma model for industrial breakdowns) *Suppose that the length of time* T *(in months) between two major breakdowns in a manufacturing company can be modeled by a gamma distribution with* $\gamma = 9$ *and* $\lambda = 0.25$*. Find the mean* μ *and the standard deviation* σ *of the time between the two major breakdowns in that company. If the model above is valid, how likely is it that the time between two major breakdowns could be as great as 45 months?*

Solution: Using the expression for the mean and the variance in Equation (5.9.17), we have

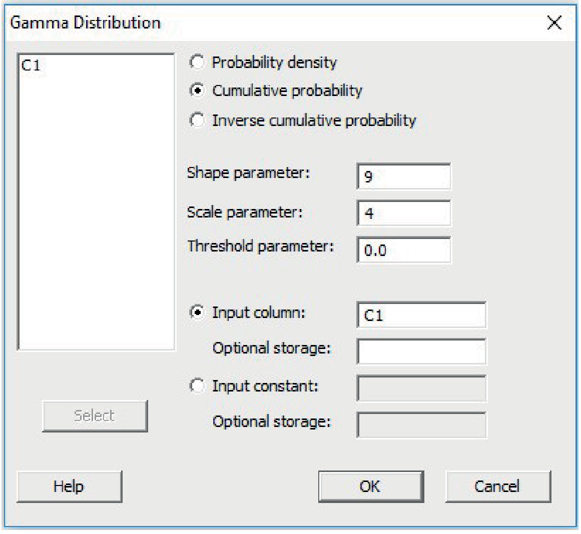
$$\mu = 9/0.25 = 36, \quad \sigma = \sqrt{\gamma}/\lambda = 3/0.25 = 12$$

Since the period of 45 months falls within one standard deviation of the mean, it is very likely that the time between the two major breakdowns could be as great as 45 months.

Example 5.9.6 (Using MINITAB and R for gamma probabilities) *The lead time (in days) for orders by company A of a certain part from a manufacturer is modeled by a gamma distribution with shape parameter* $\gamma = 9$ *and scale parameter* $1/\lambda = 4$*. Using MINITAB and R, find the probability that the lead time for an order is less than or equal to 45 days.*

MINITAB

1. Enter the value 45 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **Gamma**.
3. In the dialog box, click the circle next to **Cumulative probability**.



Gamma Distribution

C1

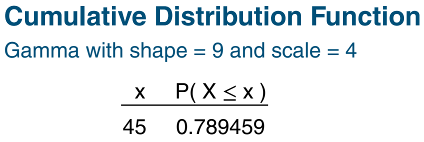
○ Probability density
● Cumulative probability
○ Inverse cumulative probability

Shape parameter: 9
Scale parameter: 4
Threshold parameter: 0.0

● Input column: C1
Optional storage:
○ Input constant:
Optional storage:

4. Enter 9 in the box next to **Shape parameter**, 4 in the box next to **Scale**, and 0 in the box next to **Threshold parameter**.

5. Click the circle next to **Input column** and type C1 in the box and Click **OK**.
6. In the session window, text appears as shown in the following box. Thus, $P(X \leq 45) = 0.789459$.



Cumulative Distribution Function

Gamma with shape = 9 and scale = 4

x	P(X ≤ x)
45	0.789459

USING R

R has a built-in cumulative gamma distribution function 'pgamma(q, shape, rate, scale = 1/rate)', where q is the quantile and shape and scale are the shape and scale parameters of the gamma distribution, respectively. Alternatively, one can specify the rate parameter, which is equal to 1/scale. So, referring to Example 5.9.6, we are asked to find $P(X \leq 45)$, which can be computed by typing the following in the R Console window.

```
pgamma(45, shape = 9, scale = 4)

#R output
[1] 0.7894591
```

This is the same value found using the MINITAB calculation.

5.9.4 The Weibull Distribution

Consider a random variable X distributed as the exponential distribution with $\lambda = 1$, so that its p.d.f. is given by $f(x) = e^{-x}, x > 0$. Now define a new random variable T, such that $X = [(T-\tau)/\alpha]^{\beta}$, $T > \tau$, so that $dx = (\beta/\alpha)[(t-\tau)/\alpha]^{\beta-1}dt$. Here, $\alpha > 0, \beta > 0$, and $\tau \geq 0$ are called the scale, shape, and threshold parameters, respectively. The resulting distribution of the random variable T is called the *Weibull distribution*. The Weibull distribution, named after its inventor, is used extensively in quality and reliability engineering (see Weibull, 1951). Unlike the exponential distribution, the Weibull distribution takes into account an aging/antiaging factor.

Definition 5.9.4 A random variable T is said to be distributed as the *Weibull distribution* if its probability density function is given by

$$f(t|\alpha,\beta,\tau) = \begin{cases} \dfrac{\beta}{\alpha}\left(\dfrac{t-\tau}{\alpha}\right)^{\beta-1} e^{-[(t-\tau)/\alpha]^{\beta}}, & \text{for} \quad t > \tau \\ 0, & \text{otherwise} \end{cases} \tag{5.9.18}$$

where $\alpha > 0, \beta > 0, \tau \geq 0$.

Usually, α is called the *scale parameter*, β the *shape parameter*, and τ the *location* or *threshold parameter*. The c.d.f. of the Weibull distribution is given by

$$F(t) = \begin{cases} P(T \leq t) = 1 - e^{-[(t-\tau)/\alpha]^{\beta}}, & \text{for} \quad t > \tau \\ 0, & \text{for} \quad t < \tau \end{cases} \tag{5.9.19}$$

From (5.9.19), it follows that

$$P(T \geq t) = 1 - P(T \leq t) = 1 - (1 - e^{-[(t-\tau)/\alpha]^{\beta}}) = e^{-[(t-\tau)/\alpha]^{\beta}} \tag{5.9.20}$$

Setting the location parameter value $\tau = 0$ and the scale parameter $\alpha = 1$ gives the Weibull distribution in its *standard* form, that is,

$$f(t|\beta) = \beta t^{\beta-1} e^{-t^{\beta}}, \quad \beta > 0, t > 0 \tag{5.9.21}$$

Graphs of the Weibull distribution for $\alpha = 1$ and $\beta = 1, 3$, and 5 for $\tau = 0$ are shown in Figure 5.9.4.

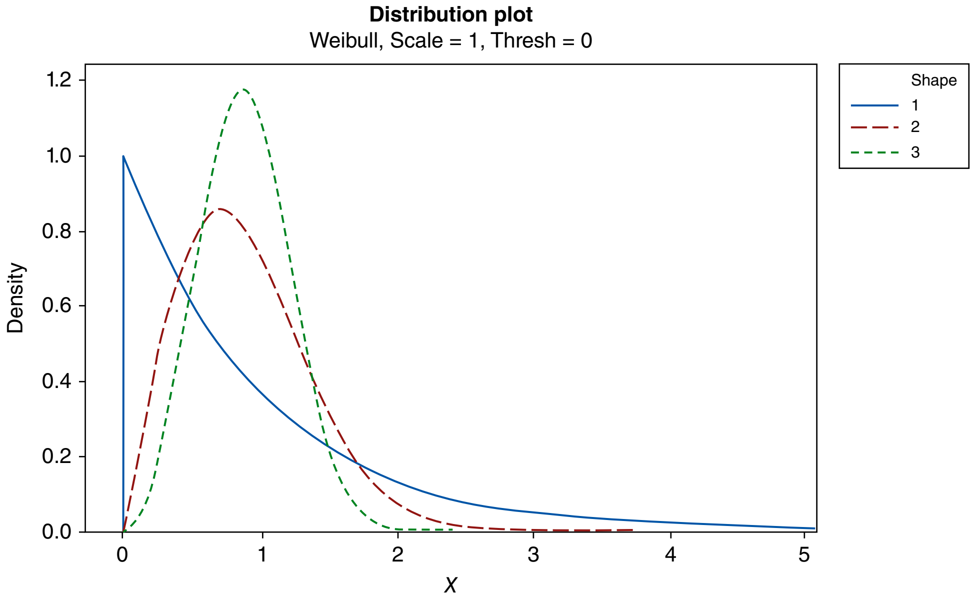


Figure 5.9.4 The p.d.f. of the Weibull distribution for various values of β and $\alpha = 1$, with $\tau = 0$.

Mean and Variance of the Weibull Distribution

Now, if T has p.d.f. (5.9.18) for $\tau = 0$, then $E(T) = \mu$ and $Var(T) = \sigma^2$ are given by

Mean and variance of Weibull distribution:

$$\mu = E(T) = \alpha\Gamma\left(1 + \frac{1}{\beta}\right) \tag{5.9.22}$$

$$\sigma^2 = Var(T) = \alpha^2 \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \left(\Gamma\left(1 + \frac{1}{\beta}\right)\right)^2 \right] \tag{5.9.23}$$

where $\Gamma(n)$ is the gamma function. As mentioned in Section 5.9.3, when n is a positive integer, then $\Gamma(n) = (n-1)!$.

Example 5.9.7 (Weibull probabilities) *From a given data set on lifetime* T *of a certain system, the parameters of a Weibull distribution are estimated to be $\alpha = 4$ and $\beta = 0.5$, where* t *is measured in thousands of hours with $\tau = 0$. Determine the following: (a) mean time before the system breaks down, (b) $P(T \geq 5)$, (c) $P(T \geq 10)$, (d) $P(T \leq 10)$.*

Solution: (a). Using the expression in (5.9.22) for the mean, we have

$$\mu = 4 \times \Gamma(1 + 1/0.5) = 4 \times \Gamma(3) = 4 \times (2!) = 8$$

(b) Using the result in (5.9.20) and taking $\tau = 0$, we have

$$P(T \geq 5) = e^{-(5/4)^{0.5}} = e^{-(1.25)^{0.5}} = e^{-1.118} = 0.3269$$

(c) Again, using the result in (5.9.20) and taking $\tau = 0$, we have

$$P(T \geq 10) = e^{-(10/4)^{0.5}} = e^{-1.581} = 0.2058$$

(d) The probability $P(T \leq 10)$ can be found using the result in (c), that is,

$$P(T \leq 10) = 1 - P(T \geq 10) = 1 - 0.2058 = 0.7942.$$

It turns out that statistical packages are able to handle problems of this type. We illustrate with the following examples.

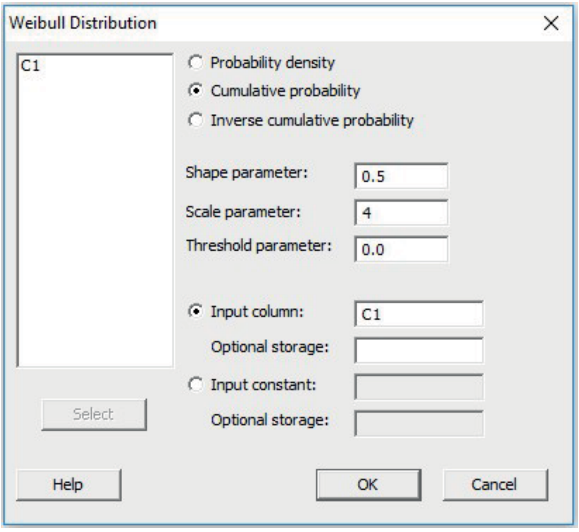
Example 5.9.8 (Using MINITAB and R for Weibull probabilities) *From a data set on a system, the parameters of a Weibull distribution are estimated to be $\hat{\alpha} = 4$ and $\hat{\beta} = 0.5$; here, time* T *is measured in thousands of hours. Using MINITAB and R, find the probability $P(T \leq 10)$.*

Solution: In order to determine the probability $P(T \leq 10)$, we proceed as follows:

MINITAB

1. Enter the value 10 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **Weibull**.

3. In the dialog box, click the circle next to **Cumulative probability**.



Weibull Distribution

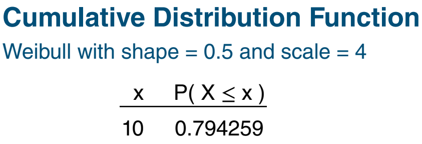
C1

○ Probability density
● Cumulative probability
○ Inverse cumulative probability

Shape parameter: 0.5
Scale parameter: 4
Threshold parameter: 0.0

● Input column: C1
Optional storage:
○ Input constant:
Optional storage:

4. Enter 0.5 in the box next to **Shape parameter**, 4 in the box next to **Scale**, and 0 in the box next to **Threshold parameter**.
5. Click the circle next to **Input column** and type C1 in the next box and Click **OK**.
6. In the session window, text appears as shown in the following box.



Cumulative Distribution Function

Weibull with shape = 0.5 and scale = 4

```
 x   P( X ≤ x )
10     0.794259
```

USING R

R has a built in cumulative Weibull distribution function ‘pweibull(q, shape, scale)’, where q is the quantile and shape and scale are the shape and scale parameters of the Weibull distribution. So, referring to Example 5.9.8, we are asked to find $P(T \leq 10)$, which can be computed by typing the following in the R Console window as follows.

```
pweibull(10, shape = 0.5, scale = 4)

#R output
[1] 0.7942593
```

This is the same value found using the MINITAB calculation.

PRACTICE PROBLEMS FOR SECTIONS 5.9.3 AND 5.9.4

1. Suppose that a random variable T denoting the time, in years, to failure of a type of computer is distributed by a gamma probability distribution with $\gamma = 6, \lambda = 2$. Find the probability that 12 of 15 such computers are still functioning after five years.

2. The lifetime, in years, of a computer hard drive follows a gamma distribution with mean of five years and standard deviation three years. Find the probability that the hard drive will last:
   (a) no more than six years, (b) at least four years, (c) between five and seven years.
3. The time, in hours, required to perform a quadruple bypass surgery follows the gamma distribution with mean eight hours and standard deviation two hours. Find the probability that a quadruple bypass surgery will last:
   (a) no more than 12 hours, (b) at least six hours, (c) between six and 10 hours.
4. Suppose that a random variable T is distributed as Weibull with parameters $\alpha = 250, \beta = 0.25$, and threshold parameter $\tau = 0$.
   (a) Find the mean and the variance of the random variable T.
   (b) Find the probability $P(5000 < T < 7000)$.
5. The lifetime, in years, of a sport utility vehicle (SUV) is distributed as Weibull with parameters $\alpha = 6, \beta = 0.5$, and threshold parameter $\tau = 0$.
   (a) Find the mean and the variance of the lifetime of the SUV.
   (b) Find the probability that the SUV will last more than 10 years.
   (c) Find the probability that the SUV will last less than 15 years.
6. The lifetime T of certain pieces of medical equipment is distributed as Weibull with parameters $\alpha = 6, \beta = 0.25$, and threshold parameter $\tau = 0$.
   (a) Find the mean and the variance of the lifetime of the equipment.
   (b) Find the probability $P(130 < T < 160)$.
7. Suppose that the time T (in hours) needed to repair a water pump can be modeled as gamma with $\gamma = 2, \lambda = 2$. What is the probability that the next repair of water pump will need:
   (a) at least 1 hour, (b) at most 2 hours, and (c) between 1 and 2 hours?
8. Suppose that the lifetime (in years) of a battery for a Movado watch is a random variable having a gamma distribution with $\gamma = 2, \lambda = 0.5$.
   (a) What is the expected life of such a battery?
   (b) What is the probability that such a battery will be working after five years?
9. Suppose that in Problem 8 the lifetime of the battery has the Weibull distribution with $\alpha = 2, \beta = 0.5$ instead of the gamma distribution of Problem 8.
   (a) What is the expected life of such a battery?
   (b) What is the probability that such a battery will be working after five years?
10. Suppose that the lifetime T (yr) of a shock absorber in a car has the Weibull distribution with $\alpha = 1, \beta = 0.4$.
    (a) What is the expected life of such a shock absorber?
    (b) What is the probability that such a shock absorber will be working after 10 years?

5.10 A CASE STUDY

Case Study (*Emissions from cars*)[1] McDonald, Vance, and Gibbons studied the effect of odometer mileage on the emission control system in cars. The data in Table 5.10.1 give

[1] Source: McDonald, Vance, and Gibbons (1995) [Data used with permission].

Table 5.10.1 Data on emissions from cars.

	0 Miles			4000 Miles			24,000 Miles			24,000 Miles		
							(before maintenance)			(after maintenance)		
Car	HC	CO	NOx	HC	CO	NOx	HC	CO	NOx	HC	CO	NOx
1	0.16	2.89	2.21	0.26	1.16	1.99	0.23	2.64	2.18	0.36	3.15	1.79
2	0.38	2.17	1.75	0.48	1.75	1.90	0.41	2.43	1.59	0.40	3.74	1.81
3	0.20	1.56	1.11	0.40	1.64	1.89	0.35	2.20	1.99	0.26	2.36	1.68
4	0.18	3.49	2.55	0.38	1.54	2.45	0.26	1.88	2.29	0.26	2.47	2.58
5	0.33	3.10	1.79	0.31	1.45	1.54	0.43	2.58	1.95	0.35	3.81	1.92
6	0.34	1.61	1.88	0.49	2.59	2.01	0.48	4.08	2.21	0.65	4.88	2.22
7	0.27	1.14	2.20	0.25	1.39	1.95	0.41	2.49	2.51	0.40	2.82	2.51
8	0.30	2.50	2.46	0.23	1.26	2.17	0.36	2.23	1.88	0.30	2.79	2.07
9	0.41	2.22	1.77	0.39	2.72	1.93	0.41	4.76	2.48	0.45	3.59	2.87
10	0.31	2.33	2.60	0.21	2.23	2.58	0.26	3.73	2.70	0.30	3.78	2.68
11	0.15	2.68	2.12	0.22	3.94	2.12	0.58	2.48	2.32	0.52	3.94	2.61
12	0.36	1.63	2.34	0.45	1.88	1.80	0.70	3.10	2.18	0.60	3.41	2.23
13	0.33	1.58	1.76	0.39	1.49	1.46	0.48	2.64	1.69	0.44	2.44	1.76
14	0.19	1.54	2.07	0.36	1.81	1.89	0.33	2.99	2.35	0.31	2.97	2.37
15	0.23	1.75	1.59	0.44	2.90	1.85	0.48	3.04	1.79	0.44	3.90	1.71
16	0.16	1.47	2.25	0.22	1.16	2.21	0.45	3.78	2.03	0.47	2.42	2.04

Note: Hydrocarbon (HC), Carbon Monoxide (CO), and Nitrogen Oxide (NOx).

three types of emissions: hydrocarbon (HC), carbon monoxide (CO), and nitrogen oxide (NOx), with their corresponding mileage conditions.

(a) Transform the CO emission at different mileage using the transformation $y = \ln x$.
(b) Plot the transformed data in each case on normal probability paper and determine whether the transformed data fits a normal model satisfactorily.
(c) If, in (b), the normal model has been found satisfactory, then use the normal probability plot in (b) to estimate the mean μ and the standard deviation σ of the normal model.
(d) Find the probability that the CO emission at 4000 miles is between 2 and 4 g.
(e) Repeat (a) through (c) for HC and NOx emissions.
(f) Find the probability that the HC and NOx emission at 4000 miles is between $(0.3, 0.7)$ and $(1.5, 3.0)$ g, respectively.

Note that in (d) through (f), it is important to use the information that if in (b) the transformed data fits a normal model satisfactorily, then the original emission values are distributed as lognormal.

5.11 USING JMP

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. A random variable X has the distribution $N(1500, (200)^2)$. Find:
   (a) $P(X < 1400)$
   (b) $P(X > 1700)$
   (c) A, where $P(X > A) = 0.05$
   (d) B, where $P(1500 - B < X < 1500 + B) = 0.95$

2. If X has the distribution $N(15.0, 6.25)$, find
   (a) $P(X < 12.0)$
   (b) $P(X > 16.5)$
   (c) C, where $P(X < C) = 0.90$
   (d) D, where $P(X > D) = 0.025$
   (e) E, where $P(|X - 15.0| < E) = 0.99$

3. Suppose that a machine set for filling 1-lb boxes of sugar yields a population of fillings, whose weights are (approximately) normally distributed with a mean of 16.30 oz and a standard deviation of 0.15 oz. Estimate:
   (a) The percentage of fillings that will be underweight (i.e., less than 1 lb)
   (b) The percentage of fillings within 16.3 ± 0.2 oz

4. Show that $P(a < X < a + l)$, where l is a positive constant and X has the distribution $N(\mu, \sigma^2)$ is maximized if $a = \mu - l/2$.

5. A process for making quarter inch ball bearings yields a population of ball bearings with diameters having mean 0.2497 in. and standard deviation of 0.0002 in. If we assume approximate normality of diameters and if specifications call for bearings with diameters to be within 0.2500 ± 0.0003 inch:
   (a) What fraction of bearings turned out under the setup are defective, that is, do not meet diameter specifications?
   (b) If minor adjustments of the process result in changing the mean diameter but not the standard deviation, what mean should be aimed at in the process setup so as to minimize the percentage of defectives?
   (c) What is the percentage of defectives in such a setup?

6. In Example 5.6.1, if "acceptable fits" are those in which the difference between hole diameter and pin diameter lies within 0.0010 ± 0.0005 inch, what fraction of random matches would yield acceptable fits?

7. Suppose that the life spans (in months) of 15 patients after they are diagnosed with a particular kind of cancer are found to be as follows:

47	50	37	44	37	44	38	35	40	38	49	42	39	38	44

   Using MINITAB, R, or JMP, verify if it is appropriate to assume that these data come from a normal population.

8. A fair coin is tossed 20 times. Find the exact and the normal approximation of the probability of obtaining 13 heads. Compare the two probabilities.

9. Suppose that 20% of the articles produced by a machine are defective, the defectives occurring at random during production. Using the normal distribution for determining approximations,
   (a) What is the approximate probability that if a sample of 400 items is taken from the production, more than 100 will be defective?
   (b) For what value of K is the probability 0.95 that the number of defectives in the sample will fall within $80 \pm K$?

10. If a sack of 400 nickels is emptied on a table and spread out, determine:
   (a) The probability (approximately) of getting between 175 and 225 heads (inclusive)?
   (b) The probability (approximately) that the number of heads is less than y? Express the answer in terms of the function $\Phi(z)$.

11. If 10% of the articles produced by a given process are defective:
   (a) What is the probability (approximately) that more than 15% of a random sample of 400 items will be defective?
   (b) For what value of K is the probability (approximately) 0.90 that the number of defectives in a sample of 400 lies within $40 \pm K$?

12. It is known that the probability of dealing a bridge hand with at least one ace is approximately 0.7. If a person plays 100 hands of bridge, what is the approximate probability
   (a) That the number of hands he/she receives containing at least one ace is between 60 and 80 inclusive?
   (b) That he/she receives at most 20 hands with no aces?

13. A die is rolled 720 times. Using a normal distribution to approximate probabilities, estimate the probability that
   (a) More than 130 sixes turn up.
   (b) The number of sixes obtained lie between 100 and 140 inclusive.

14. A mass-produced laminated item is made up of five layers. A study of the thickness of individual layers shows that each of the two outside layers have mean thickness of 0.062 in., and each of the three middle layers have mean thickness 0.042 in. The standard deviation of thickness of outside layers is 0.004 in. and that of inside layers is 0.003 in.
   (a) If random assembly is employed, what are the mean thicknesses of the laminated items?
   (b) What is the standard deviation of the thicknesses of the laminated items?
   (c) Assuming the thicknesses of the individual sections to be approximately normally distributed, what percentage of items have thicknesses between 0.240 and 0.260 in.?
   (d) For what value of K will 90% of the items have thicknesses falling within $0.250 \pm K$?

15. An article is made up of three independent parts A, B, and C. The weights of the A's have an (approximately) normal distribution with mean 2.05 oz and standard deviation 0.03 oz. Those of the B's have an (approximately) normal distribution with mean 3.10 oz and standard deviation 0.04 oz. Those of the C's have an (approximately) normal distribution with mean 10.5 oz and deviation of 0.12 oz. Then approximately
   (a) What fraction of the assembled articles have weights exceeding 1 lb?

(b) The probability is 0.95 that four articles picked at random will have a total weight less than what value?

16. Suppose that items of a certain kind are counted by weighing a box of 100 of these items. The population of individual items has a mean weight of 1.45 oz and standard deviation of 0.02 oz. A batch of items weighing between 144.5 and 145.5 items is counted as 100 items. What is the probability:

    (a) That a batch of 100 items will be counted correctly?
    (b) That 101 items will be passed as 100?

17. A resistor is composed of eight component parts soldered together in series, so that the total resistance of the resistor equals the sum of the resistances of the component parts. Three of the components are drawn from a production lot that has a mean of 200 Ω and a standard deviation of 2 Ω, four components from a lot that has a mean of 150 Ω and standard deviation of 3 Ω, and one component from a lot that has a mean of 250 Ω and standard deviation of 1 Ω. Assume that the total resistance of the assembled resistors is approximately normally distributed.

    (a) Five percent of such resistors have a resistance less than what value?
    (b) What is the probability that a sample of four such resistors manufactured from these components have an average resistance in excess of 1443 Ω?

18. Let Z be a random variable distributed as the standard normal. Determine the probability that the random variable Z takes a value (a) within one standard deviation of the mean, (b) within two standard deviation of the mean, (c) within three standard deviation of the mean.

19. Let Z be a random variable distributed as the standard normal. Using the normal distribution table (Table A.4) determine the following probabilities: (a) $P(Z \leq 2.11)$, (b) $P(Z \geq -1.2)$, (c) $P(-1.58 \leq Z \leq 2.40)$, (d) $P(Z \geq 1.96)$, (e) $P(Z \leq -1.96)$.

20. Let X be a random variable distributed by the exponential distribution with $\lambda = 1.5$. Determine the probability that the random variable X assumes a value: (a) greater than 2, (b) less than 4, (c) between 2 and 4, (d) less than 0.

21. Let X be a random variable distributed as an exponential distribution with $\lambda = 2$. Determine the probability that the random variable X assumes a value: (a) greater than 1, (b) greater than 2, (c) between 1 and 2, (d) greater than 0.

22. Let X be a random variable distributed as the Weibull distribution with $\alpha = 100$ and $\beta = 0.5$. Determine the mean and the variance of the random variable X. (Assume that τ, the threshold parameter, has value $\tau = 0$.)

23. In Problem 22, determine the probability that the random variable X assumes a value: (a) greater than 450 and (b) greater than 700.

24. Suppose that the life of a motor (in hours) follows the Weibull distribution with $\alpha = 1000$ and $\beta = 2.0$. Determine the mean and the variance of the random variable X.

25. In Problem 24, determine the probabilities of the following events:

    (a) The motor fails before 800 hours.
    (b) The motor lasts more than 1000 hours.
    (c) The motor lasts between 1000 and 1500 hours.

26. The time (in hours) needed to finish a paint job of a car is an exponentially distributed random variable with $\lambda = 0.2$.
    (a) Find the probability that a paint job exceeds seven hours.
    (b) Find the probability that a paint job exceeds seven hours but finishes before 10 hours.
    (c) Find the probability that a paint job lasts at least eight hours, given that it exceeds five hours.
    (d) Find the probability that a paint job finishes before seven hours.

27. The number of years a computer functions is exponentially distributed with $\lambda = 0.1$. David bought a five-year-old computer that is functioning well. What is the probability that David's computer will last another nine years?

28. Suppose that the lifetime of a serpentine belt of a car is distributed as an exponential random variable with $\lambda = 0.00125$. What is the probability that a serpentine belt lasts
    (a) 700 hours?
    (b) More than 850 hours?
    (c) Between 600 and 900 hours?
    (d) At least 650 hours?

29. Suppose that the length of time X (in months) taken by two different medications, say Lipitor and Zocor, to lower the bad cholesterol (LDL) level by 20 mg/dl can be modeled by two gamma distributions with parameters $\gamma = 3, \lambda = 1$ and $\gamma = 6, \lambda = 1.5$, respectively.
    (a) Find the mean and variance of time taken by the medication Lipitor to lower the bad cholesterol (LDL) level by 20 mg/dl.
    (b) Find the mean and variance of time taken by the medication Zocor to lower the bad cholesterol (LDL) level by 20 mg/dl.

30. In Problem 29, find the following probabilities:
    (a) Lipitor takes at least two months to lower the bad cholesterol (LDL) level by 20 m·g/dl
    (b) Zocor takes at least three months to lower the bad cholesterol (LDL) level by 20 mg/dl

31. If X is a continuous random variable with p.d.f. $f(x)$, the pth percentile of x (sometimes called the$p$th percentile of the population) is defined as that value x_p for which

$$P(X \leq x_p) = \int_{-\infty}^{x_p} f(x)dx = p/100$$

The 50th percentile is called the *population median.* Also, suppose that a continuous random variable has cumulative distribution

$$F(x) = \begin{cases} 0, & x > 1 \\ x^n, & 0 < x \leq 1 \\ 1, & x > 1 \end{cases}$$

where $n \geq 1$. Then,
    (a) Find the probability density function $f(x)$.
    (b) Find the median of X.
    (c) Find the mean and variance of X.

32. If a defective spot occurs in a glass disk R inches in radius, assume that it is equally likely to occur anywhere on the disk. Let X be a random variable indicating the distance between the point of occurrence of a defective spot and the center of the disk.
    (a) Find the expression for $F(x)$ and $f(x)$.
    (b) Find the median of X.
    (c) Find the mean and variance of the random variable X.

33. A continuous random variable X has the probability density function

$$f(x) = \begin{cases} 3x^2, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

    (a) Find the c.d.f. $F(x)$.
    (b) Find the numerical values of $F(1/3), F(9/10)$, and $P(1/3 < X \leq 1/2)$.
    (c) Find the value of a which is such that $P(X \leq a) = 1/4$ (a is the 25th percentile of X).
    (d) Find the mean μ and variance σ^2 of X.

34. Determine which of the following functions are probability density functions:
    (a) $f(x) = x(3-x),\ 0 \leq x \leq 3$
    (b) $f(x) = x^2(3-x),\ 0 \leq x \leq 3$
    (c) $f(x) = x(3-x),\ 0 \leq x \leq 2$
    (d) $f(x) = \dfrac{1}{\lambda}e^{-(x-2)/\lambda},\ x \geq 2$

35. Determine the value of c so that the following functions are probability density functions:
    (a) $f(x) = cx(3-x),\ 0 \leq x \leq 3$
    (b) $f(x) = cx^2(3-x),\ 0 \leq x \leq 3$
    (c) $f(x) = cx^3(3-x),\ 0 \leq x \leq 1$

36. Find the mean and the variance for each of the following probability density functions:
    (a) $f(x) = 1/2,\ 1 \leq x \leq 3$, and zero elsewhere.
    (b) $f(x) = \theta e^{-\theta x},\ x > 0, \theta > 0$, and zero elsewhere.
    (c) $f(x) = 12x^2(1-x),\ 0 \leq x \leq 1$, and zero elsewhere.

37. Referring to Problem 36, in (a) and (c) find the exact probability $P(|X-\mu| \leq 2\sigma)$. Then, find the lower bound of these probabilities using Chebyshev's inequality. Compare the two results and comment.

38. Repeat Problem 37 for the exact probability $P(|X-\mu| \leq 4\sigma)$.

39. Suppose that the length of life (in months) of a computer chip follows a gamma distribution with $\gamma = 4, \lambda = 0.05$. Determine the following probabilities:
    (a) $P(40 < X < 120)$
    (b) $P(X > 80)$
    (c) $P(X < 100)$

40. If the life (in thousands of miles) of a car tire follows a gamma distribution with $\gamma = 6, \lambda = 0.1$, determine the following probabilities:
    (a) $P(35 < X < 85)$
    (b) $P(X > 75)$
    (c) $P(X < 50)$

41. In Problem 39, find the mean and the variance of X, the length of life of the computer chip.

42. Referring to Problem 40, suppose that a customer bought a new car with tires of that brand. Find the probabilities of the following events:
    (a) All the tires will last at least 50,000 miles.
    (b) At least one tire will last 50,000 miles or more.
    (c) None of the tires will last more than 50,000 miles.

43. The time between arrivals (in minutes) of customers at a teller's window in a bank, is a gamma random variable with $\gamma = 1, \lambda = 0.1$. Find the following probabilities:
    (a) The time between arrivals of two customers is more than 10 minutes.
    (b) The time between arrivals of two customers is less than 15 minutes.
    (c) The time between arrivals of two customers is more than 10 minutes, but less than 15 minutes.

44. The waiting time, say X hours in an emergency room, is distributed as a gamma with mean $\mu = 2$ and variance $\sigma^2 = 3$.
    (a) Determine the probability density function for the waiting time.
    (b) Determine the probability that randomly selected patient has to wait more than 2.5 hours.

45. Referring to Problem 44:
    (a) Use Chebyshev's theorem to find an interval that contains at least 88.8% of the waiting times.
    (b) Determine the actual probability of waiting times to fall in the interval you determined in part (a).

46. The time lapse between two accidents in a large manufacturing plant has an approximately exponential distribution with a mean of two months.
    (a) What is the probability that the time lapse between two accidents is less than three months?
    (b) What is the probability that the time lapse between two accidents is less than two months?

47. Referring to Problem 46:
    (a) Determine the probability that at least two accidents take place in three months.
    (b) Determine the probability that less than two accidents take place in two months.

48. Suppose that a random variable X (in thousands) is distributed as lognormal with parameters $\mu = 3$ and $\sigma^2 = 4$. Determine the following probabilities: (a) $P(X \leq 5500)$, (b) $P(X \geq 2000)$.

49. Suppose that random variable X (in thousands) is distributed as lognormal with parameters $\mu = 5$ and $\sigma^2 = 9$. Determine the following probabilities: (a) $P(3500 \leq X \leq 9500)$, (b) $P(1500 \leq X \leq 2500)$.

50. Suppose that random variable X (in hundreds) is distributed as lognormal with parameters $\mu = 2$ and $\sigma^2 = 4$. Determine the following probabilities: (a) $P(X \leq 750)$, (b) $P(X \geq 1500)$.

51. Suppose that random variable X is distributed as lognormal with parameters $\mu = 2$ and $\sigma^2 = 4$ Find the mean and variance of X.

52. Suppose that random variable X is distributed as normal with parameters $\mu = 2$ and $\sigma^2 = 4$. Find the value of x such that (a) $P(X \leq x) = 0.05$, (b) $P(X \geq x) = 0.33$.

53. The life (in hours) of a domestic dehumidifier is modeled by a Weibull distribution with parameters $\alpha = 300$ and $\beta = 0.25$ hour. Assuming $\tau = 0$,
    (a) What is the mean life of the dehumidifier?
    (b) What is the variance of the life of the dehumidifier?

54. Referring to Problem 53:
    (a) What is the probability that the dehumidifier fails before 5000 hours?
    (b) What is the probability that the dehumidifier lasts between 8000 to 12,000 hours?
    (c) What is the probability that the dehumidifier lasts at least 7000 hours?

55. The life (in hours) of a catalytic converter of a passenger car is modeled by a Weibull distribution with parameters $\alpha = 2000$ and $\beta = 0.4$. (Assume $\tau = 0$).
    (a) What is the probability that the catalytic converter needs to be replaced before 14,000 hours?
    (b) What is the probability that the catalytic converter lasts between 12,000 to 16,000 hours?
    (c) What is the probability that the catalytic converter lasts at least 16,000 hours?

56. The time needed (in hours) for a worker to finish a job is modeled by a lognormal distribution with parameters $\mu = 2$ and $\sigma^2 = 4$. Find the following probabilities:
    (a) The worker needs at least 50 hours to finish the job.
    (b) The worker needs more than 60 hours to finish the job.
    (c) The worker needs less than 50 hours to finish the job.

57. Suppose that a random variable X is distributed as an exponential with mean 20. Find the following probabilities:
    (a) $P(X > 25)$
    (b) $P(15 < X < 25)$
    (c) $P(X \geq 20)$

58. The time (in units of 500 hours) between the two consecutive breakdowns of a machine is modeled by an exponential distribution with mean of 1.2.
    (a) What is the probability that the second breakdown does not occur for at least 600 hours after the first breakdown? [*Note*: $600 = (1.2) \times 500$.]
    (b) What is the probability that the second breakdown occurs less than 400 hours after the first breakdown?

59. Referring to Problem 58:
    (a) What is the probability that more than two breakdowns occur in 1000 hours?
    (b) What is the probability that at least two breakdowns occur in 1000 hours?
    (c) What is the probability that less than two breakdowns occur in 1000 hours?

60. The time between arrivals of cars in a garage is exponentially distributed with a mean time between arrivals of cars of 30 minutes.
    (a) What is the probability that the time between arrivals of two successive cars is more than 45 minutes?
    (b) What is the probability that the time between arrivals of two successive cars is less than 20 minutes?
    (c) What is the probability that two cars arrive within a 40-minute interval?

61. Suppose that a random variable X is distributed as binomial with $n = 225$ and $\theta = 0.2$. Using the normal approximation, find the following probabilities: (a) $P(X \leq 60)$, (b) $P(X \geq 57)$, (c) $P(80 \leq X \leq 100)$.

62. The fluid volume of a popular drink in a 12-oz can is normally distributed with mean 12.3 oz and standard deviation 0.2 oz.
    (a) What is the probability that a randomly selected can has less than 12 oz of drink?
    (b) What is the probability that a randomly selected can has more than 12.5 oz of drink?

63. A random variable X is distributed uniformly over the interval $[0, 20]$. Determine the probabilities: (a) $P(X < 5)$, (b) $P(3 < X < 16)$, (c) $P(X > 12)$.

64. Suppose that X is the length of a rod that is uniformly distributed over the specification limits 19 and 20 cm. Find the probabilities: (a) $P(19.2 < X < 19.5)$, (b) $P(X < 19.5)$, (c) $P(X > 19.7)$.

65. Referring to Problem 64:
    (a) Find the mean and the variance of the random variable X
    (b) What percentage of the rods is within two standard deviations of the mean?

66. Assume that the log of failure times are normally distributed, with parameters μ and σ^2. A sample of 10 parts selected at random has failure times whose logs are

7.77	8.45	7.59	7.03	7.17	6.46	7.46	9.09	7.81	7.47

Use normal probability paper to determine the approximate values of μ and σ^2.

67. Suppose that time (in hours) to failure of a machine is modeled by the lognormal with parameters μ and σ^2. The failure time of 12 such machines are as follows:

269	207	214	254	739	580	267	725	154	306	439	215

Estimate the values of μ and σ^2.

68. Refer to Problem 67. Determine the probability that the time to failure of a machine is: (a) less than 200 hours, (b) between 300 and 500 hours, (c) more than 600 hours

69. The following data give time T to failure of a drug in suspension form:

5	7	9	13	24	32	35	38	42	47	49	52

The lifetime of the suspension follows the lognormal distribution with parameters μ and σ^2. Estimate the values of μ and σ^2.

Chapter 6

DISTRIBUTION OF FUNCTIONS OF RANDOM VARIABLES

The focus of this chapter is on the distributions of functions of random variables.

Topics Covered

- Joint distributions of two discrete random variables
- Joint distributions of two continuous random variables
- Mean value and variance of functions of two random variables
- Conditional distributions
- Correlation between two random variables
- Joint distributions of several random variables
- Moment-generating functions of two or more random variables

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the concept of discrete and continuous distributions of functions of two or more random variables and use them to solve real-world problems.
- Understand the concept of marginal and conditional probability distributions.
- Determine the moment-generating functions of functions of two or more random variables.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

6.1 INTRODUCTION

In Chapters 4 and 5, we discussed various phenomena by enlisting a (single) random variable and studying its distribution and characteristics. However, in many situations, experiments are performed that involve two or more random variables. For example, we may be interested in the diameter and length of rods, the number of dots on two dice when rolled simultaneously, say (X, Y), where $1 \leq X \leq 6, \quad 1 \leq Y \leq 6$, or the composition of a Monel (70% nickel, 30% copper) alloy, where we may focus on solid contents, say X, and liquid content Y, which again we would quote as a joint pair (X, Y).

In this chapter, then, we will study the joint distribution functions of two or more discrete and continuous random variables.

6.2 DISTRIBUTION FUNCTIONS OF TWO RANDOM VARIABLES

6.2.1 Case of Two Discrete Random Variables

If, for each element e in a finite sample space S, we make two measurements on e, say $(X(e),\ Y(e))$, and if (x_i, y_j), $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, are possible values of $(X(e),\ Y(e))$, and if we let

$$p_{ij} = p(x_i, y_j) = P(X(e) = x_i, Y(e) = y_j) \tag{6.2.1}$$

then the set of all possible values $\{(x_i, y_j)\}$ of $(X(e),\ Y(e))$ is called the *sample space* of $(X(e),\ Y(e))$, while the set of associated probabilities p_{ij} is the joint probability function (p.f.) of the pair of discrete random variables (X, Y).

Thus, we may think of $k = mn$ points (x_i, y_j) in the xy-plane in which the probabilities p_{ij} are located and are all positive and sum to 1. If we define $p_{i.}$ and $p_{.j}$ such that

$$p_{i.} = \sum_{j=1}^{n} p_{ij} \quad \text{and} \quad p_{.j} = \sum_{i=1}^{m} p_{ij} \tag{6.2.2}$$

then

$$p_{i.} = P(X(e) = x_i) \quad \text{and} \quad p_{.j} = P(Y(e) = y_j) \tag{6.2.3}$$

The possible values x_i, $i = 1, 2, \ldots, m$, of $X(e)$ together with their probabilities $p_{i.}$ constitute the *marginal distribution* of the random variable X. This gives rise to the probability function of X, ignoring Y, and is therefore merely the probability function of X. In a similar manner, the y_j, $j = 1, 2, \ldots, n$, of $Y(e)$ together with their probabilities $p_{.j}$ constitute the *marginal distribution* of the random variable Y.

Geometrically, if x is the usual horizontal axis and y the vertical axis and if we project the sum of the probabilities $p_{i1}, \ldots, p_{ij}, \ldots, p_{in}$ located at the points $[(x_i, y_1), \ldots, (x_i, y_j), \ldots, (x_i, y_n)]$, vertically onto the x-axis, we obtain the marginal distribution $p_{i.}$ of the random variable X. If instead we project sum of these probabilities $p_{1j}, \ldots, p_{ij}, \ldots, p_{mj}$ horizontally onto the y-axis, we obtain the marginal distribution $p_{.j}$ of the random variable Y.

The mean μ_1 and variance σ_1^2 of X are defined by applying (4.2.1) and (4.2.2) to the probability distribution $p_{i.}$. Similarly, the mean μ_2 and variance σ_2^2 of Y are defined by applying those formulas to $p_{.j}$.

When the probability function p_{ij} factors into the product of the two marginal probability functions, that is, if for all possible (x_i, y_j) in the sample space of (X, Y), we have

$$p_{ij} = p_{i.}p_{.j} \tag{6.2.4}$$

then X and Y are said to be *independent random variables.*

Example 6.2.1 (Probability function of two random variables) *Roll a pair of fair dice, of which one die is green and the other is red. Let the random variables X and Y denote the outcomes on the green and red dies, respectively. Then, the sample space of (X, Y) is $S = \{(1, 1), (1, 2), \ldots, (1, 6), \ldots, (6, 6)\}$. Each of the 36 sample points has the probability 1/36. Then, the* joint probability function *of the random variables X and Y can be written in tabular form as follows:*

X Y	1	2	3	4	5	6	Total $(p_{.j})$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	$\vdots$	$\vdots$		$\vdots$	$\vdots$	1/6
3	1/36	$\cdots$	1/36	$\cdots$	1/36	1/36	1/6
4	1/36	$\vdots$	$\vdots$		$\vdots$	$\vdots$	1/6
5	1/36	$\cdots$	1/36	$\cdots$	1/36	1/36	1/6
6	1/36	$\cdots$	1/36	$\cdots$	1/36	1/36	1/6
Total $(p_{i.})$	1/6	1/6	1/6	1/6	1/6	1/6	1

This table shows the probabilities assigned to each sample point. Using (6.2.2) for the probabilities, we easily find the marginal distributions $p_{i.}$ and $p_{.j}$ of the random variables X and Y, respectively, as shown in the table. The probability function in this example can also be expressed as

$$p(x, y) = \begin{cases} 1/36, & x = 1, 2, 3, 4, 5, 6 \text{ and } y = 1, 2, 3, 4, 5, 6 \\ 0, & \text{otherwise} \end{cases}$$

We give a graphical representation of the p.f. of X and Y in Figure 6.2.1.

Example 6.2.2 (Marginal probability functions) *Let the joint probability function of random variables X and Y be defined as*

$$P(x, y) = \frac{x + y}{54}, \quad x = 1, 2, 3, 4; \quad y = 1, 2, 3$$

Find the marginal probability functions of X *and* Y *and also examine whether X and Y are independent.*

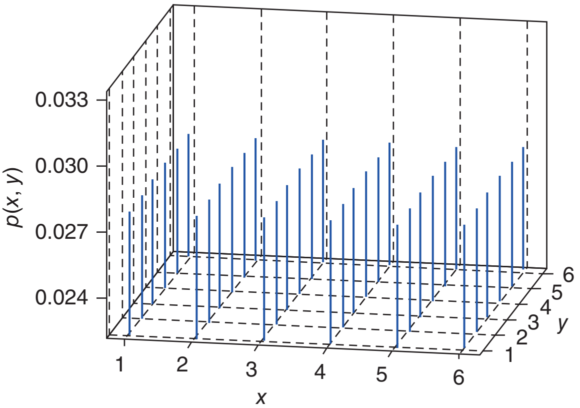


Figure 6.2.1 Graphical representation of the p.f. in Example 6.2.1.

Solution: From equation (6.2.2), it follows that the probability of X, say $p_1(x)$, is given by

$$P(X = x) = p_1(x) = \sum_{y=1}^{3} \frac{x+y}{54} = \frac{x+1}{54} + \frac{x+2}{54} + \frac{x+3}{54} = \frac{3x+6}{54}, \quad \text{for} \quad x = 1, 2, 3, 4$$

Similarly,

$$P(Y = y) = p_2(y) = \sum_{x=1}^{4} \frac{x+y}{54} = \frac{10+4y}{54}, \quad \text{for } y = 1, 2, 3$$

For (x, y) belonging to the sample space of (X, Y), say $S_{xy} = \{(x, y)|x = 1, 2, 3, 4; y = 1, 2, 3\}$, we have that

$$p(x, y) \neq p_1(x) \times p_2(y)$$

so that the random variables X and Y are not independent.

Example 6.2.3 (Joint probability function and its marginals) *In dealing a hand of 13 cards from a deck of ordinary playing cards, let X_1 and X_2 be random variables denoting the numbers of spades and of hearts, respectively. Obviously, $0 \leq X_1 \leq 13$, $0 \leq X_2 \leq 13$, and $0 \leq X_1 + X_2 \leq 13$. Then, we see that $p(x_1, x_2)$, the p.f. of (x_1, x_2), is given by*

$$p(x_1, x_2) = \frac{\binom{13}{x_1}\binom{13}{x_2}\binom{26}{13-x_1-x_2}}{\binom{52}{13}}$$

where the sample space of (X_1, X_2) is all pairs of nonnegative integers (x_1, x_2) for which $0 \leq x_1, x_2 \leq 13$ and $0 \leq x_1 + x_2 \leq 13$. That is, the sample space $\{(x_1, x_2)\}$ consists of the 105 points:

$$\{(0, 0), \ldots, (0, 13), \ldots, (12, 0), (12, 1), (13, 0)\}$$

Now, it is possible by a direct probability argument to find the marginal distribution of X_1, for the probability of x_1 spades in a hand of 13 is clearly given by

$$P(X_1 = x_1) = p_1(x_1) = \frac{\binom{13}{x_1}\binom{39}{13-x_1}}{\binom{52}{13}}$$

where $0 \leq x_1 \leq 13$.

In a similar manner, it is easy to find $p_2(x_2)$ and to show that the random variables X_1 and X_2 are not independent.

6.2.2 Case of Two Continuous Random Variables

If the sample space S consists of a continuum of elements and if for any point (x_1, x_2) in the x_1x_2-plane we let

$$F(x_1, x_2) = P[X_1(e) \leq x_1, X_2(e) \leq x_2] \tag{6.2.5}$$

then $F(x_1, x_2)$ is called the *cumulative distribution function* (c.d.f.) of the pair of random variables (X_1, X_2) (dropping e). If there exists a nonnegative function $f(x_1, x_2)$ such that

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2) dt_2 dt_1 \tag{6.2.6}$$

then

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$$

and $f(x_1, x_2)$ is called the *joint probability density function* (p.d.f.) of the pair of random variables (X_1, X_2). The probability that this pair of random variables represents a point in a region E, that is, the probability that the event E occurs, is given by

$$P((X_1, X_2) \in E) = \iint_E f(x_1, x_2) dx_2 dx_1 \tag{6.2.7}$$

Note that if $E = \{(X_1, X_2)|X_1 < x_1, X_2 < x_2\}$, then (6.2.7) equals $F(x_1, x_2)$. Also, if we let

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \tag{6.2.8}$$

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 \tag{6.2.9}$$

then $f_1(x_1)$ and $f_2(x_2)$ are called the *marginal probability density functions* of X_1 and X_2, respectively. This means that $f_1(x_1)$ is the p.d.f. of X_1 (ignoring X_2), and $f_2(x_2)$ is the p.d.f. of X_2 (ignoring X_1).

Geometrically, if we think of $f(x_1, x_2)$ as a function describing the manner in which the total probability 1 is continuously "smeared" in the x_1x_2-plane, then the integral in (6.2.7) represents the amount of probability contained in the region E. Also, $f_1(x_1)$ is the p.d.f. one obtains by projecting the probability density in the x_1x_2-plane orthogonally onto the x_1-axis, and $f_2(x_2)$ is similarly obtained by orthogonal projection of the probability density onto the x_2-axis.

If $f(x_1, x_2)$ factors into the product of the two marginal p.d.f.'s, that is, if

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) \tag{6.2.10}$$

for all (x_1, x_2) in the sample space of (X_1, X_2), then X_1 and X_2 are said to be *independent continuous random variables.*

Example 6.2.4 (Marginal probability functions) *Let the joint probability density function of the random variables X_1 and X_2 be defined as*

$$f(x_1, x_2) = 2e^{-(2x_1+x_2)}; \quad x_1 > 0, \quad x_2 > 0$$

Find the marginal probability density functions of X_1 and X_2 and examine whether or not X_1 and X_2 are independent.

Solution: From equations (6.2.8) and (6.2.9), it follows that for $x_1 > 0$

$$\begin{aligned} f_1(x_1) &= \int_0^\infty 2e^{-(2x_1+x_2)}dx_2 = 2e^{-2x_1}\int_0^\infty e^{-x_2}dx_2 \\ &= 2e^{-2x_1}[-e^{-x_2}]_0^\infty = 2e^{-2x_1}, \quad x_1 > 0 \end{aligned}$$

while for $x_2 > 0$,

$$\begin{aligned} f_2(x_2) &= \int_0^\infty 2e^{-(2x_1+x_2)}dx_1 = e^{-x_2}\int_0^\infty 2e^{-2x_1}dx_1 \\ &= e^{-x_2}\left[\frac{2e^{-2x_1}}{-2}\right]_0^\infty = e^{-x_2}, \quad x_2 > 0 \end{aligned}$$

Here, we clearly have that $f(x_1, x_2) = f_1(x_1)f_2(x_2)$, which implies that the random variables X_1 and X_2 are independent.

Finally, note that the joint distribution function satisfies the properties given below:

1. $0 \le F(x_1, x_2) \le 1$ for all (x_1, x_2) belong to the sample space of (X_1, X_2).
2. $F(-\infty, x_2) = F(x_1, -\infty) = F(-\infty, \infty) = F(\infty, -\infty) = 0, F(\infty, \infty) = 1$.
3. F is nondecreasing.
4. For every pair of (X_1, X_2) values, say x_{i1} and x_{i2} where $x_{i1} < x_{i2}$ for $i = 1, 2$, the following inequality holds:

$$F(x_{12}, x_{22}) - F(x_{12}, x_{21}) - F(x_{11}, x_{22}) + F(x_{11}, x_{21}) \ge 0 \tag{6.2.11}$$

The reader should verify that the left-hand side of (6.2.11) gives the value of $P(x_{11} < X_1 < x_{12}, x_{21} < X_2 < x_{22})$.

6.2.3 The Mean Value and Variance of Functions of Two Random Variables

Suppose that (X_1, X_2) is a pair of discrete random variables and $g(X_1, X_2)$ is a function of (X_1, X_2). Then, the mean value or expectation of $g(X_1, X_2)$, say $E(g(X_1, X_2))$, is given by

$$E(g(X_1, X_2)) = \sum\sum g(x_{1i}, x_{2j})p(x_{1i}, x_{2j}) \tag{6.2.12}$$

where the summation $\sum\sum$ is over all pairs (x_{1i}, x_{2j}) in the sample space of (X_1, X_2), and for the continuous case we have that

$$E(g(X_1, X_2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f(x_1, x_2) dx_1 dx_2 \tag{6.2.13}$$

We may now state, and the reader should verify equation (6.2.14) in Theorem 6.2.1 stated below.

Theorem 6.2.1 *If X_1 and X_2 are independent random variables and if $g_1(X_1)$ and $g_2(X_2)$ depend only on X_1 and X_2, respectively, then*

$$E(g_1(X_1)\ g_2(X_2)) = E(g_1(X_1))\ E(g_2(X_2)) \tag{6.2.14}$$

If we choose $g(X_1, X_2)$ as $(X_1 - \mu_1)(X_2 - \mu_2)$, we obtain the *covariance*, which is a measure of the relationship between two random variables X_1 and X_2, that is,

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] \tag{6.2.15}$$

In the case where X_1 and X_2 are independent, we find that

$$Cov(X_1, X_2) = E(X_1 - \mu_1)E(X_2 - \mu_2) = 0 \tag{6.2.16}$$

In many problems, however, we deal with linear functions of two or even more independent random variables. The following theorem is of particular importance in this connection:

Theorem 6.2.2 *Let X_1 and X_2 be independent random variables such that the mean and variance of X_1 are μ_1 and σ_1^2, and the mean and variance of X_2 are μ_2 and σ_2^2. Then, if c_1 and c_2 are constants, $c_1X_1 + c_2X_2$ is a random variable having mean value $c_1\mu_1 + c_2\mu_2$ and variance $c_1^2\sigma_1^2 + c_2^2\sigma_2^2$.*

Proof: To prove this theorem, it is sufficient to consider the case of continuous random variables. (The proof for discrete random variables is obtained by replacing integral signs by signs of summation.) For the mean value of $c_1X_1 + c_2X_2$, we have, since X_1 and X_2 are independent, that

$$\begin{aligned}
E(c_1X_1 + c_2X_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (c_1x_1 + c_2x_2) f_1(x_1) f_2(x_2) dx_1 dx_2 \\
&= c_1 \int_{-\infty}^{\infty} x_1 f_1(x_1) dx_1 \int_{-\infty}^{\infty} f_2(x_2) dx_2 + c_2 \int_{-\infty}^{\infty} x_2 f_2(x_2) dx_2 \int_{-\infty}^{\infty} f_1(x_1) dx_1 \\
&= c_1 E(X_1) + c_2 E(X_2) \\
&= c_1\mu_1 + c_2\mu_2
\end{aligned}$$

For the variance of $c_1X_1 + c_2X_2$, we have similarly (omitting some straightforward details), that

$$\begin{aligned} Var(c_1X_1 + c_2X_2) &= E[c_1(X_1 - \mu_1) + c_2(X_2 - \mu_2)]^2 \\ &= c_1^2E(X_1 - \mu_1)^2 + c_2^2E(X_2 - \mu_2)^2 + 2c_1c_2E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= c_1^2E(X_1 - \mu_1)^2 + c_2^2E(X_2 - \mu_2)^2 + 0 \\ &= c_1^2\sigma_1^2 + c_2^2\sigma_2^2 \end{aligned}$$

since X_1 and X_2 are independent, so that $E[(X_1 - \mu_1)(X_2 - \mu_2)] = 0$, as stated in (6.2.16). We remark that if X_1, X_2 are not independent (and the reader should verify), then we have that

$$\begin{aligned} E(c_1X_1 + c_2X_2) &= c_1\mu_1 + c_2\mu_2 \\ Var(c_1X_1 + c_2X_2) &= c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + 2c_1c_2Cov(X_1, X_2) \end{aligned}$$

□

In a straightforward manner, it is easy to prove the following theorem, which extends the results of this section:

Theorem 6.2.3 *Let $X_1, X_2, \ldots, X_n$ be* n *random variables such that the mean and variance of X_i are μ_i and σ_i^2 respectively, and where the covariance of X_i and X_j is σ_{ij}, that is, $E[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ij}$, $i \neq j$. If $c_1, c_2, \ldots, c_n$ are constants, then the random variable $L = c_1X_1 + \cdots + c_nX_n$ has mean value and variance that are given by*

$$E(L) = c_1\mu_1 + \cdots + c_n\mu_n \tag{6.2.17}$$

$$Var(L) = c_1^2\sigma_1^2 + \cdots + c_n^2\sigma_n^2 + 2c_1c_2\sigma_{12} + 2c_1c_3\sigma_{13} + \cdots + 2c_{n-1}c_n\sigma_{n-1,n} \tag{6.2.18}$$

Further, if $X_1, X_2, \ldots, X_n$ are mutually independent, then $\sigma_{ij} = 0$, so that the mean of L is as in (6.2.17). However, the variance of L is

$$Var(L) = c_1^2\sigma_1^2 + \cdots + c_n^2\sigma_n^2 \tag{6.2.19}$$

6.2.4 Conditional Distributions

Suppose that a pair of discrete random variables (X_1, X_2) has joint p.f. $p(x_1, x_2)$ and marginal probability functions $p_1(x_1)$ and $p_2(x_2)$, as defined in Section 6.2.1. Suppose that we assign to one of the random variables, say X_1, a value x_1 such that $p_1(x_1) \neq 0$, and we want to find the probability that the other random variable X_2 has a particular value, say x_2. The required probability is a *conditional probability* that we may denote by $p(X_2 = x_2|X_1 = x_1)$, or, more briefly, by $p(x_2|x_1)$, and is defined as follows:

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p_1(x_1)} \tag{6.2.20}$$

where $p_1(x_1) \neq 0$.

Note that $p(x_2|x_1)$ has all the properties of an ordinary probability function; that is, as the reader should verify, the sum of $p(x_2|x_1)$ over all possible values of x_2, for fixed x_1, is 1. Thus, $p(x_2|x_1)$, $x_2 = x_{21}, \ldots, x_{2k_2}$, is a p.f. and is called the *conditional probability function of* X_2*, given that* $X_1 = x_1$.

Note that we can write (6.2.20) as

$$p(x_1, x_2) = p_1(x_1) \cdot p(x_2|x_1) \tag{6.2.21}$$

to provide a two-step procedure for finding $p(x_1, x_2)$ by first determining $p_1(x_1)$, then $p(x_2|x_1)$, and by multiplying the two together.

Example 6.2.5 (Conditional probability function) *In Example 6.2.3, suppose that we want to find the conditional probability function of* X_2 *given* $X_1 = x_1$*, that is,* $p(x_2|x_1)$.

Solution: The probability function of X_1 is given by

$$p_1(x_1) = \frac{\binom{13}{x_1}\binom{39}{13-x_1}}{\binom{52}{13}}$$

Hence, as is easily verified, $p(x_2|x_1)$ is given by

$$p(x_2|x_1) = \frac{\binom{13}{x_2}\binom{26}{13-x_1-x_2}}{\binom{39}{13-x_1}}$$

where the sample space of X_2, given $X_1 = x_1$, is $\{0, 1, \ldots, 13 - x_1\}$. The interpretation of $p(x_2|x_1)$ is that if we are given that a hand of 13 cards contains x_1 spades, then the value of $p(x_2|x_1)$ as given previously is the probability that the hand also contains $X_2 = x_2$ hearts.

In the case of a pair of continuous random variables (X_1, X_2) having probability density function $f(x_1, x_2)$ and marginal probability density functions $f_1(x_1)$ and $f_2(x_2)$, the conditional probability density function $f(x_2|x_1)$ of X_2 given $X_1 = x_1$ is defined as

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} \tag{6.2.22}$$

where $f_1(x_1) \neq 0$, which is the analogue of (6.2.20), now for a pair of continuous random variables. Note that $f(x_2|x_1)$ has all the properties of an ordinary probability density function.

Now from (6.2.22), we have the result that is given below.

$$f(x_1, x_2) = f_1(x_1)\ f(x_2|x_1) \tag{6.2.22a}$$

We now have the analogue of (6.2.21) for obtaining the probability density function of a pair of continuous random variables in two steps, as given in (6.2.22a).

Example 6.2.6 (Determination of marginal probability functions) *Suppose that we are dealing with a pair of continuous random variables* (X_1, X_2) *whose sample space* S *is given by* $S = \{(x_1, x_2)|0 \leq x_1, x_2 \leq 1; 0 \leq x_1 + x_2 \leq 1\}$. *Suppose further that the probability density function* $f(x_1, x_2)$ *of* (X_1, X_2) *is given by*

$$f(x_1, x_2) = \begin{cases} 2, & \text{if} \quad (x_1, x_2) \in S \\ 0, & \text{otherwise} \end{cases}$$

Solution: Because the probability density function is constant over the triangle defined by S in the x_1, x_2-plane (see Figure 6.2.2), we sometimes say that (X_1, X_2) is *uniformly* distributed over S.

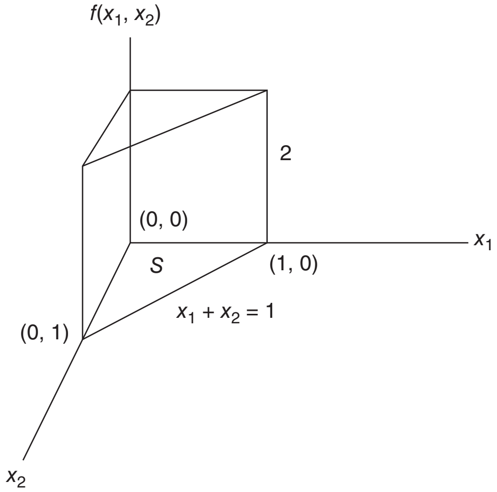


Figure 6.2.2 Graphical representation of the p.d.f. in Example 6.2.6.

The marginal probability density function of X_1 for $0 < x_1 < 1$ is given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = 2 \int_0^{1-x_1} dx_2 = 2(1 - x_1)$$

Hence,

$$f(x_2|x_1) = \begin{cases} \frac{2}{2(1-x_1)} = \frac{1}{(1-x_1)}, & \text{for} \quad 0 < x_2 < 1 - x_1 \\ 0, & \text{otherwise} \end{cases}$$

Note that if (X_1, X_2) is a pair of discrete random variables, then the conditional mean and variance of X_2 given $X_1 = x_1$ are defined as given at (6.2.23) and (6.2.24).

$$E(X_2|X_1 = x_1) = \sum_{x_2} x_2 p(x_2|x_1) \tag{6.2.23}$$

$$Var(X_2|X_1 = x_1) = \sum_{x_2} [x_2 - E(x_2|x_1)]^2 p(x_2|x_1) \tag{6.2.24}$$

where $p(x_2|x_1)$ is the conditional probability function of the random variable X_2 given $X_1 = x_1$. The mean and variance for other functions of X_2 given $X_1 = x_1$ can be defined in the same manner.

Similarly, for the case of a pair of continuous random variables, we have the following results.

$$E(X_2|X_1 = x_1) = \int_{-\infty}^{\infty} x_2 f(x_2|x_1) dx_2 \tag{6.2.25}$$

$$Var(X_2|X_1 = x_1) = \int_{-\infty}^{\infty} [x_2 - E(x_2|x_1)]^2 f(x_2|x_1) dx_2 \tag{6.2.26}$$

6.2.5 Correlation between Two Random Variables

The reader will note from (6.2.15) that the covariance between the random variables X_1 and X_2 is a quantity measured in [(units of X_1) × (units of X_2)]. A somewhat more convenient measure of how X_1 and X_2 "co-vary," or are dependent on each other, is the *theoretical* or *population correlation coefficient* ρ. This dimensionless measure of dependence is defined by

$$\rho = Cov(X_1, X_2)/\sigma_1\sigma_2 \tag{6.2.27}$$

where σ_1 and σ_2 are the population standard deviations of X_1 and X_2, respectively. It can be shown that $-1 \leq \rho \leq 1$, and hence, we have that $-\sigma_1\sigma_2 \leq Cov(X_1, X_2) \leq \sigma_1\sigma_2$.

Now, from (6.2.16) and using (6.2.17), we have that if X_1 and X_2 are independent random variables, then $\rho = 0$. The converse need not be true however, as the following example shows.

Example 6.2.7 (*Independence and correlation coefficient*) *Two random variables X_1 and X_2 have joint probability function given by*

$$p(x_1, x_2) = \begin{cases} 1/3, & \text{if} \quad , (x_1, x_2) = (0,0), (1,1), (2,0) \\ 0, & \text{otherwise} \end{cases}$$

It is easy to see that

$$p_1(x_1) = \begin{cases} 1/3, & \text{if} \quad x_1 = 0, 1, 2 \\ 0, & \text{otherwise} \end{cases}$$

and that

$$p_2(x_2) = \begin{cases} 2/3, & \text{if} \quad x_2 = 0 \\ 1/3, & \text{if} \quad x_2 = 1 \end{cases}$$

Hence, $p(0,0) \neq p_1(0)p_2(0)$, and so on, and X_1 and X_2 are *not independent*. Simple calculations further show that

$$\mu_1 = E(X_1) = 1, \quad \mu_2 = E(X_2) = 1/3$$

$$\sigma_1^2 = E(X_1 - \mu_1)^2 = 2/3, \quad \sigma_2^2 = E(X_2 - \mu_2)^2 = 2/9$$

Also, since $(X_1 - \mu_1)(X_2 - \mu_2) = (X_1 - 1)(X_2 - 1/3)$, we have that

$$\begin{aligned} Cov(X_1, X_2) &= \sum\sum(x_1 - 1)(x_2 - 1/3)p(x_1, x_2) \\ &= 1/3[(0-1)(0-1/3) + (1-1)(1-1/3) + (2-1)(0-1/3)] \\ &= 1/3[1/3 + 0 - 1/3] \\ &= 0 \end{aligned}$$

Therefore, the correlation coefficient has value $\rho = 0$, yet X_1 and X_2 are not independent.

Example 6.2.8 (Joint probability density function) *Let the length of life (in years) of both an operating system and the hard drive of a computer be denoted by the random variables X_1 and X_2, respectively. Suppose that the joint distribution of the random variables of X_1 and X_2 is given by*

$$f(x_1, x_2) = \begin{cases} \frac{1}{64}x_1^2x_2e^{-(x_1+x_2)/2}, & \text{if} \quad x_1 > 0, \quad x_2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

The probability density function is graphed in Figure 6.2.3.

(a) Find the marginal distributions of the random variables X_1 and X_2.
(b) Find the mean and variance of the random variables X_1 and X_2.
(c) Examine whether the random variables X_1 and X_2 are independent.
(d) Find $Cov(X_1, X_2)$.

Solution:
(a) The marginal probability density function of X is given by

$$f_1(x_1) = \int_0^\infty f(x_1, x_2)dx_2 = \int_0^\infty \frac{1}{64}x_1^2x_2e^{-(x_1+x_2)/2}dx_2 = \frac{1}{64}x_1^2e^{-x_1/2}\int_0^\infty x_2e^{-x_2/2}dx_2$$

Integrating by parts, we have that

$$\begin{aligned} f_1(x_1) &= \frac{1}{64}x_1^2e^{-x_1/2}\left[x_2\frac{e^{-x_2/2}}{-1/2}|_0^\infty - \int_0^\infty 1 \times \frac{e^{-x_2/2}}{-1/2}dx_2\right] \\ &= \frac{1}{32}x_1^2e^{-x_1/2}\int_0^\infty e^{-x_2/2}dx_2 \\ &= \frac{1}{16}x_1^2e^{-x_1/2}, \quad x_1 > 0 \end{aligned}$$

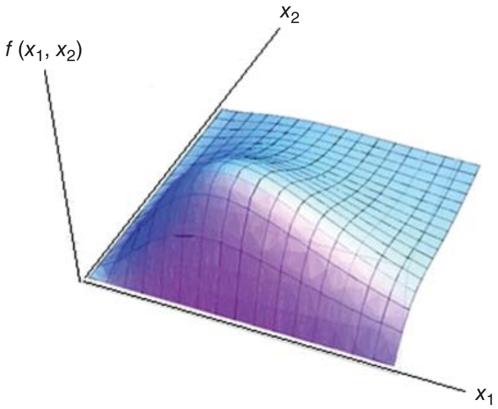


Figure 6.2.3 Graphical representation of the joint p.d.f. in Example 6.2.8.

Similarly, it can be shown (as the reader should verify) that

$$f_2(x_2) = \frac{1}{4}x_2 e^{-x_2/2}, \quad x_2 > 0$$

Comparing the marginal distribution of X_1 and X_2 with the gamma distribution given in equation (5.9.10), we can see that the random variables X_1 and X_2 are distributed as gamma with parameters $\gamma = 3, \lambda = 1/2$, and $\gamma = 2, \lambda = 1/2$, respectively.

(b) Since the random variables X_1 and X_2 are distributed marginally as gamma, using equation (5.9.11), we have

$$\mu_1 = 6, \quad \sigma_1^2 = 12 \quad \text{and} \quad \mu_2 = 4, \quad \sigma_2^2 = 8$$

(c) We also have that

$$f_1(x_1) \times f_2(x_2) = \frac{1}{16}x_1^2 e^{-x_1/2} \times \frac{1}{4}x_2^2 e^{-x_2/2} = \frac{1}{64}x_1^2 x_2 e^{-(x_1+x_2)/2} = f(x_1, x_2)$$

so that the random variables X_1 and X_2 are independent.

(d) Since the random variables X_1 and X_2 are independent, we have that $Cov(X_1, X_2) = 0$.

We now state an important result about the expected value of sum of functions of random variables X_1 and X_2, similar to the one for a single variable.

Theorem 6.2.4 *Let X_1 and X_2 be random variables and $g_i(X_1, X_2)$, $i = 1, 2, \ldots, m$ be m functions of X_1 and X_2. Then,*

$$E\left[\sum_{i=1}^{m} g_i(X_1, X_2)\right] = \sum_{i=1}^{m} E(g_i(X_1, X_2)) \tag{6.2.28}$$

This result can be proved in exactly the same manner as in single random-variable (univariate) case.

Theorem 6.2.5 *Let X_1 and X_2 be random variables with means μ_1 and μ_2, respectively. Then,*

$$Cov(X_1, X_2) = E(X_1X_2) - \mu_1\mu_2 \tag{6.2.29}$$

From equation (6.2.15), we have

$$\begin{aligned} Cov(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1X_2 - X_1\mu_2 - \mu_1X_2 + \mu_1\mu_2] \end{aligned}$$

Using Theorem 6.2.4 with $g_1(X_1, X_2) = X_1X_2$, $g_2(X_1, X_2) = -\mu_2X_1$, $g_3(X_1, X_2) = -\mu_1X_2$, and $g_4(X_1, X_2) = \mu_1\mu_2$, we find that

$$\begin{aligned} Cov(X_1, X_2) &= E(X_1X_2) - \mu_2E(X_1) - \mu_1E(X_2) + \mu_1\mu_2 \\ &= E(X_1X_2) - \mu_1\mu_2 - \mu_1\mu_2 + \mu_1\mu_2 \\ &= E(X_1X_2) - \mu_1\mu_2 \end{aligned}$$

The reader should now use the result of Theorem 6.2.1 together with equation (6.2.29) to show the following corollary:

Corollary 6.2.1 *If X_1 and X_2 are two independent random variables, $Cov(X_1, X_2) = 0$.*

6.2.6 Bivariate Normal Distribution

Consider a pair of continuous random variables (X_1, X_2). These random variables (X_1, X_2) are said to be distributed as the *bivariate normal* if their joint p.d.f. $f(x_1, x_2)$ is given below.

$$\begin{aligned} f(x_1, x_2) = {} & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \\ & \times \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right) \end{aligned} \tag{6.2.30}$$

where $-\infty \leq x_i \leq \infty, -\infty \leq \mu_i \leq \infty, \sigma_i^2 > 0, \quad i = 1, 2$, and $-1 < \rho < 1$.

When plotted in three dimensions, a typical bivariate normal probability density function takes the form given in Figure 6.2.4. We say that this probability density function has parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$, and ρ, and it can be shown that μ_1, μ_2 and σ_1, σ_2 are means and standard deviations of the random variables X_1 and X_2, respectively. Further, ρ is the correlation coefficient, where $-1 < \rho < 1$. By integrating $f(x_1, x_2)$ over $-\infty \leq x_1 \leq \infty$

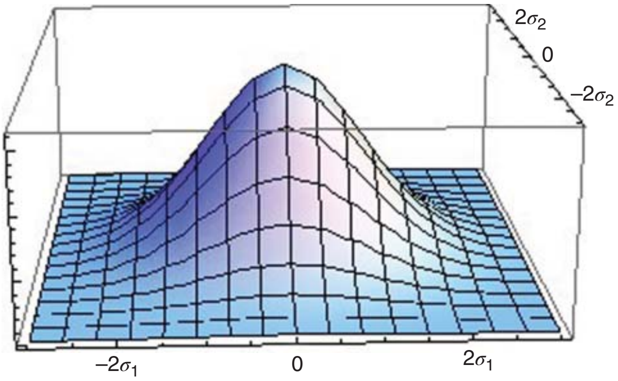


Figure 6.2.4 Graphical representation of p.d.f. of bivariate normal.

and $-\infty \leq x_2 \leq \infty$, it can be seen that the marginal distributions of random variables X_1 and X_2 are given by

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right), -\infty \leq x_1 \leq \infty \tag{6.2.31}$$

and

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right), -\infty \leq x_2 \leq \infty \tag{6.2.32}$$

respectively. Furthermore, $f(x_1, x_2)$ in (6.2.30) may be written as

$$\begin{aligned} f(x_1, x_2) = {} & \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right) \\ & \times \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_2} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right)^2\right) \end{aligned} \tag{6.2.33}$$

for $-\infty \leq x_1, x_2 \leq \infty$.

From equations (6.2.31) and (6.2.33), it follows that the conditional p.d.f. of the random variable X_2 given $X_1 = x_1$ after some algebraic manipulation is as stated below.

$$\begin{aligned} f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = {} & \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2^2}} \\ & \times \exp\left(-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(x_2 - \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)^2\right) \end{aligned} \tag{6.2.34}$$

for $-\infty \leq x_2 \leq \infty$.

The p.d.f. in (6.2.34) is a normal density function with mean $\mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1)$ and variance $\sigma_2^2(1-\rho^2)$. In a similar fashion, we can show that the conditional p.d.f. of the random variable X_1 given $X_2 = x_2$ is given by

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_1^2}} \times \exp\left(-\frac{1}{2\sigma_1^2(1-\rho^2)}\left(x_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)\right)^2\right) \qquad (6.2.35)$$

for $-\infty \leq x_1 \leq \infty$,

which is a normal density function with mean $\mu_1 + \rho(\sigma_1/\sigma_2)(x_2 - \mu_2)$ and variance $\sigma_1^2(1 - \rho^2)$.

From equations (6.2.30) through (6.2.32), it can easily be seen that $\rho = 0$ implies that X_1 and X_2 are independent. Thus, we state this important result as a corollary below.

Corollary 6.2.2 *If X_1 and X_2 are distributed as bivariate normal, then the random variables X_1 and X_2 are independent if and only if $\rho = 0$.* ***Note that this result is not true in general.***

Example 6.2.9 (Independence and correlation coefficient) *Let a random variable* X *be distributed as the uniform over an interval $(-a, a)$, and let $Y = X^b$ be another random variable where* b *is an even integer. Clearly, the random variables X and Y are not independent, but the reader can easily show that the correlation coefficient between X and Y is $\rho = 0$.*

PRACTICE PROBLEMS FOR SECTION 6.2

1. (a) *X, Y*, and *Z* are independent, Poisson random variables with mean 2, 7, and 9, respectively. Determine the mean and variance of the random variable $U = 5X + 3Y + 8Z$.
   (b) Let $(X_1, \ldots, X_5)$, $(Y_1, \ldots, Y_3)$, and $(Z_1, \ldots, Z_8)$ be random samples from the three Poisson populations with mean 2, 7, and 9, respectively. Determine the mean and variance of the random variable $U = \sum_{i=1}^{5} X_i + \sum_{j=1}^{3} Y_j + \sum_{k=1}^{8} Z_k$.
2. In filling soft drinks into 12-oz cans, assume the population of net amounts of drinks generated by the automatic filling machine, adequately calibrated, has a distribution with mean of 12.15 oz and standard deviation 0.1 oz. Assume that the population of aluminum cans used for fillings have a distribution with mean 1.5 oz and standard deviation of 0.05 oz.
   (a) The population of filled cans will have a distribution with what mean and variance?
   (b) If these cans are packed into boxes of 48 cans each and if the population of empty boxes has mean 30 oz and standard deviation 2 oz, what is the mean and variance of the population of filled boxes?
3. A sole of a running shoe is built by randomly selecting one layer of material I, one layer of material II, and two layers of material III. The thicknesses of individual

layers of material I, II, and III have distributions with means 0.20, 0.30, and 0.15 cm and standard deviations 0.02, 0.01, and 0.03 cm, respectively. Find the mean and standard deviation of thicknesses of a sole in this lot.

4. If X_1 and X_2 have joint distribution $f(x_1, x_2)$, show that $Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$. Hence show that if X_1 and X_2 are independent, then $Cov(X_1, X_2) = 0$.
5. Referring to Example 6.2.6, find the marginal distributions of X_1 and X_2. What are the mean and variances of X_1 and X_2? Compute the correlation coefficient between X_1 and X_2.
6. Referring to Example 6.2.6, find $E(X_1|X_2 = x_2)$, $Var(X_1|X_2 = x_2)$, $E(X_2|X_1 = x_1)$, and $Var(X_2|X_1 = x_1)$.
7. Referring to Example 6.2.7, find conditional probability functions $p(x_1|x_2)$ and $p(x_2|x_1)$, and evaluate $E(X_1|X_2 = 0)$, $Var(X_1|X_2 = 0)$, $E(X_1|X_2 = 1)$, $Var(X_1|X_2 = 1)$.
8. The pair of random variables (X_1, X_2) has joint p.d.f. $f(x_1, x_2)$ given by $f(x_1, x_2) = 2/\pi$ for (x_1, x_2) lying inside the semicircle bounded by the x_1 axis and the curve $x_2 = \sqrt{1 - x_1^2}$; that is, the sample space of (X_1, X_2) is $S = \{(x_1, x_2)|x_1^2 + x_2^2 \leq 1, x_2 \geq 0\}$. Find the marginals of X_1 and X_2, the means and variances of X_1 and X_2, and the correlation coefficient between X_1 and X_2. Also determine $f(x_1|x_2)$ and $f(x_2|x_1)$, and evaluate $E(X_1|X_2 = x_2)$, $Var(X_1|X_2)$, and $E(X_2|X_1)$, $Var(X_2|X_1)$.
9. During rush hours, an engineer takes X minutes to reach to his/her office. This time includes (Y) the driving time from home to the parking lot and walking time from parking lot to the office. Thus, $U = X - Y$ is the time that he/she has taken to find a parking spot. Suppose that the joint distribution of the random variables X and Y has joint p.d.f. $f(x, y)$ given by

$$f(x, y) = \begin{cases} e^{-x}, & \text{if} \quad 0 \leq y < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

Determine the density function of the waiting time $U = X - Y$.

6.3 EXTENSION TO SEVERAL RANDOM VARIABLES

The notions of (i) the probability function of a pair of discrete random variables and (ii) the probability density function of a pair of continuous random variables extend without special difficulties to sets of three or more random variables. Consider then the case of n discrete random variables $(X_1, \ldots, X_n)$. Suppose that the sample space of $(X_1, \ldots, X_n)$ is the set of k possible points

$$(x_{11}, \ldots, x_{n1}), \ldots, (x_{1k}, \ldots, x_{nk})$$

in an n-dimensional space with associated probabilities

$$p(x_{11}, \ldots, x_{n1}), \ldots, p(x_{1k}, \ldots, x_{nk})$$

respectively, which are all positive and whose sum is 1. If all the probability is projected orthogonally onto the x_1-axis, we obtain the marginal probability function of X_1, say

$p_1(x_1)$. Similarly, we obtain marginal probability functions $p_2(x_2), \ldots, p_n(x_n)$ of the remaining random variables $X_2, \ldots, X_n$. We now have the following important result, given below.

> If $p(x_1, \ldots, x_n)$ factors into the product of the marginal probability functions, that is, if
>
> $$p(x_1, \ldots, x_n) = p_1(x_1) \cdots \; p_n(x_n) \tag{6.3.1}$$
>
> we say that the random variables $X_1, \ldots, X_n$ are *mutually independent.*

In the case of n continuous random variables, $X_1, \ldots, X_n$, suppose that we have a probability density function $f(x_1, \ldots, x_n)$ that is nonnegative throughout the entire n-dimensional space of the variables. The probability that $X_1, \ldots, X_n$ falls into any region or set E (i.e., the probability that event E occurs) is given by

$$P[(x_1, \ldots, x_n) \in E] = \int \cdots \int_E f(x_1, \ldots, x_n) dx_1 \; \cdots dx_n$$

By setting

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_2 \; \cdots dx_n$$

with similar definitions for $f_2(x_2), \ldots, f_n(x_n)$, we obtain the marginal probability density functions $f_1(x_1), \ldots, f_n(x_n)$ of $X_1, \ldots, X_n$, respectively.

> If $f(x_1, \ldots, x_n)$ factors as
>
> $$f(x_1, \ldots, x_n) = f_1(x_1) \cdots \; f_n(x_n), \tag{6.3.2}$$
>
> then the random variables $X_1, \ldots, X_n$ are said to be *mutually independent* random variables.

The extensions of Theorems 6.2.1 and 6.2.2 to the case of n-independent random variables are straightforward and are left to the reader. The concepts of a conditional p.f. and of a conditional p.d.f. of X_n given $X_1, \ldots, X_{n-1}$ are also straightforward and left to the reader.

6.4 THE MOMENT-GENERATING FUNCTION REVISITED

Suppose that X_1 and X_2 are two independent random variables and c_1 and c_2 are constants. Let $M_{X_1}(t)$ and $M_{X_2}(t)$ be the moment-generating functions of X_1 and X_2, respectively. Then, the moment-generating function of the random variable

$$U = c_1X_1 + c_2X_2 \tag{6.4.1}$$

is

$$M_U(t) = E(e^{(c_1X_1+c_2X_2)t})$$

Using Theorem 6.2.1 of Section 6.2.3, we have, since X_1 and X_2 are independent, that

$$M_U(t) = E(e^{X_1c_1t})E(e^{X_2c_2t}) = M_{X_1}(c_1t)M_{X_2}(c_2t) \tag{6.4.2}$$

Indeed, if we are interested in the linear combination

$$U = \sum_{i=1}^{k} c_iX_i \tag{6.4.3}$$

where $X_i, \quad i = 1, \ldots, k$, are independent random variables, we find similarly that

$$M_U(t) = \prod_{i=1}^{k} M_{X_i}(c_it) \tag{6.4.4}$$

The moment-generating function has an important property given by the following uniqueness theorem, which we state without proof

Theorem 6.4.1 *If two random variables X and Y have the same moment-generating function $M(t)$, then their c.d.f.'s are identical.*

Example 6.4.1 (Distribution of the sum of n Poisson random variables) *Suppose that a discrete random variable X has the Poisson distribution with a parameter λ, then from (4.8.4), the moment-generating function of X is*

$$M_X(t) = e^{\lambda(e^t-1)}$$

Suppose now that $X_1, \ldots, X_n$ are n independent observations on some characteristic X. This means that each of the X_i's have the same Poisson distribution, so that $X_1, \ldots, X_n$ is a random sample from the Poisson with parameter λ, and we have that

$$P(X_i = x_i) = p(x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \tag{6.4.5}$$

Note that each X_i has the same mean, variance, and in particular, the same moment-generating function; that is,

$$M_{X_i}(t) = e^{\lambda(e^t-1)} \tag{6.4.6}$$

Suppose that we want to find the moment-generating function of

$$U = X_1 + \cdots + X_n \tag{6.4.7}$$

In the language of (6.4.3), each of the c_i is 1, and using (6.4.4), we find that

$$M_U(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda(e^t-1)} = e^{n\lambda(e^t-1)} \tag{6.4.8}$$

Thus, it follows from Theorem 6.4.1 and equation (6.4.6) that the random variable U has the Poisson distribution with "parameter" $n\lambda$; that is, we have

$$p(u) = \frac{e^{-n\lambda}(n\lambda)^u}{u!} \tag{6.4.9}$$

Example 6.4.2 (MGF of the gamma distribution) *Suppose that a random variable* V *has a gamma distribution of order* m*; that is, its probability density function* f *is given by*

$$f(v) = \begin{cases} \frac{1}{\Gamma(m)} v^{m-1} e^{-v}, & \text{if} \quad v > 0, \ m > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.4.10}$$

$\Gamma(m)$ is called the gamma function of order m, $m > 0$, and is defined as

$$\Gamma(m) = \int_0^\infty t^{m-1} e^{-t} dt, \quad m > 0 \tag{6.4.11}$$

It can be shown by integrating by parts that

$$\Gamma(m) = (m-1)\Gamma(m-1) \tag{6.4.12}$$

and that for integer m,

$$\Gamma(m) = (m-1)! \tag{6.4.13}$$

We now determine the moment-generating function of the random variable V, that is,

$$M_V(t) = \frac{1}{\Gamma(m)} \int_0^\infty v^{m-1} e^{-v(1-t)} dv \tag{6.4.14}$$

Using the transformation $w = v(1-t)$, we find that

$$\begin{aligned} M_V(t) &= (1-t)^{-m} \frac{1}{\Gamma(m)} \int_0^\infty w^{m-1} e^{-w} dw \\ &= (1-t)^{-m} \frac{1}{\Gamma(m)} \Gamma(m) \end{aligned}$$

or

$$M_V(t) = (1-t)^{-m} \tag{6.4.15}$$

Now suppose that $X_1, \ldots, X_m$ is a *random sample* from a single exponential population with $\lambda = 1$; that is, the X_i is m independent observations on a random variable X, where X has distribution given by

$$f(x) = \begin{cases} e^{-x}, & \text{if} \quad x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.4.16}$$

Then, from equation (5.9.8), the moment-generating function of X is given by

$$M_X(t) = (1-t)^{-1} \tag{6.4.17}$$

Hence, if $U = X_1 + \cdots + X_m$, then, because of the independence of the X_i's,

$$M_U(t) = \prod_{i=1}^{m} M_{X_i}(t) = \prod_{i=1}^{m} (1-t)^{-1} = (1-t)^{-m} \tag{6.4.18}$$

That is, from equation (6.4.15), we have

$$M_U(t) = M_V(t).$$

Invoking Theorem 6.4.1, we thus have that the distribution of U is that of a gamma random variable of order m, so that the form of the probability density function is as in (6.4.10), that is

$$f(u) = \begin{cases} \frac{1}{\Gamma(m)} u^{m-1} e^{-u}, & \text{if} \quad u > 0, \;\; m > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.4.19}$$

Suppose now that $X_1, \ldots, X_i, \ldots, X_n$ are *independent normal random variables* with mean μ_i and variance σ_i^2, for $i = 1, \ldots, n$, respectively, and suppose that

$$U = \sum_{i=1}^{n} c_i X_i$$

Then, using equations (5.5.11) and (6.4.4), we have

$$M_U(t) = \exp\left\{ \sum_{i=1}^{n} (c_i \mu_i) t + \sum_{i=1}^{n} (c_i^2 \sigma_i^2) t^2 \right\} \tag{6.4.20}$$

Suppose now that $(X_1, \ldots, X_n)$ is a random sample from a *normal population* with mean μ and variance σ^2, and suppose that

$$U = \frac{1}{n}(X_1 + \cdots + X_n) = \bar{X}$$

Then, again using equations (5.5.11) and (6.4.4), we have

$$M_{\bar{X}}(t) = \exp\left\{ \mu t + \frac{1}{2} \frac{\sigma^2}{n} t^2 \right\} \tag{6.4.21}$$

Now using Theorem 6.4.1, and equations (6.4.20) and (6.4.21), we have two very important results that we now state in Theorems 6.4.2 and 6.4.3, given below,

Theorem 6.4.2 *If $X_1, \ldots, X_i, \ldots, X_n$ are independent normal random variables with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$, for $i = 1, \ldots, n$, then any linear combination $U = c_1 X_1 + \cdots + c_n X_n$ of $X_1, \ldots, X_i, \ldots, X_n$ is normally distributed with mean $c_1 \mu_1 + \cdots + c_n \mu_n$ and variance $c_1^2 \sigma_1^2 + \cdots + c_n^2 \sigma_n^2$.*

Theorem 6.4.3 *If $X_1, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2, then $\bar{X}$ is also normally distributed with mean μ and variance σ^2/n.*

PRACTICE PROBLEMS FOR SECTIONS 6.3 AND 6.4

1. Let X_1 and X_2 be two independent random variables distributed as standard normal. Find the moment-generating function of the random variable $U = (X_1 - X_2)$. Find the mean and variance of U.
2. Let (X_1, X_2, X_3) be a random sample from a Poisson population with mean λ. Determine the moment-generating function of the random variable $Y = 2X_1 + 3X_2 + X_3$ and use it to find the mean and the variance of the random variable Y.
3. Let X_1 and X_2 be two independent random variables distributed as standard normal. Find the moment generating function of the random variable $V = (2X_1 + 3X_2)$. Find the mean and variance of V.
4. Suppose that $X_1, \ldots, X_n$ are independent and identically distributed random variables with probability function for X_i given by

$$P(X_i = x_i) = p(x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, \ldots, n$$

   That is, the X_i's constitute a random sample of n independent observations on X, where X has the Poisson distribution with parameter λ. Find the moment-generating function of the random variable $Y = X_1 + \cdots + X_n$.
5. Referring to Problem 9 of Section 6.2, find the moment-generating function of the random variable $U = X - Y$.

Review Practice Problems

1. Suppose that $X_1, \ldots, X_n$ are independent continuous random variables such that $X_i \quad (i = 1, 2, \ldots, n)$ is gamma distributed with parameters γ_i and λ. Find the moment-generating function of the random variable $Y = X_1 + \cdots + X_n$.

2. Suppose that $X_1, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2. Find the moment-generating function of the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$.

3. Suppose that $X_1, \ldots, X_n$ are independent random variables such that $X_i \quad (i = 1, 2, \ldots, n)$ is distributed with probability function

$$P(X_i = x_i) = p(x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, \ldots, n.$$

   Find the moment-generating function of the random variable $Y = X_1 + \cdots + X_n$.

4. Suppose that $X_1, \ldots, X_n$ are independently distributed with $N(\mu_i, \sigma_i^2), \imath = 1, 2, \ldots, n$. Find the moment-generating function of the random variable $Y = X_1 + \cdots + X_n$.

5. A resistor is composed of two component parts soldered together in series; the total resistance of the resistor equals the sum of the resistances of the component parts. The first part is drawn from a production lot having a mean of 200 Ω and standard deviation of 2 Ω, and the second part is drawn from a lot having a mean of 150 Ω and standard deviation of 3 Ω. Find the mean and standard deviation of the resistance of the assembled resistor.

6. In packaging corn flakes into 8-oz packages, assume that the population of net weights generated by the automatic filling machine (properly calibrated) has a distribution with mean of 8.15 oz and standard deviation of 0.08 oz. Assume that the population of paper boxes to receive the fillings to have a distribution with mean 1.45 oz and standard deviation of 0.06 oz.
   (a) The population of filled boxes will have a distribution with what mean and variance?
   (b) If these boxes are packaged 24 per carton and if the population of empty cartons has mean 28.00 oz and standard deviation 1.20 oz, what is the mean and variance of the population of filled cartons?

7. A laminated strip is built up by randomly selecting two layers of material A, three layers of material B, and four layers of material C. The thicknesses of the individual layers of material A have mean 0.0100 in., and standard deviation 0.0005 in.; the respective numbers for material B are 0.0050 and 0.0003 in., and those for C are 0.0025 and 0.0001 in. A large lot of such laminated strips are manufactured. Find the mean and standard deviation of the thicknesses of the strips in this lot.

8. Mass-produced articles are fitted into cardboard containers, one article to a container. Twelve of these filled containers are then packed in wooden boxes. Suppose that the mean and standard deviation of the weights in pounds (lb) of the population of articles are 20.6 and 0.8 lb respectively, those of the cardboard containers are 1.8 and 0.1 lb respectively, and those of the wooden boxes are 3.6 and 0.4 lb, respectively.
   (a) What are the values of the mean and standard deviation of the population of weights of filled boxes ready to ship?
   (b) Let T be the total weight of 25 filled wooden boxes taken at random. Find the mean and variance of T.
   (c) Suppose that $\bar{X}$ is the average weight of those 25 boxes. Find the mean and variance of $\bar{X}$.

9. A certain type of half-inch rivet is classified as acceptable by a consumer if its diameter lies between 0.4950 and 0.5050 in. It is known that a mass-production process is such that $100p_1\%$ of the rivets have diameters less than 0.4950 in., $100p_2\%$ have diameters that lie in the "acceptable" region, and $100p_3\%$ have diameters greater than 0.5050 in., where, of course, $p_3 = 1 - p_1 - p_2$. If a random sample of n rivets is taken from the process, what is the probability $p(x_1, x_2)$ that $X_1 = x_1$ rivets have diameters less than 0.4950 in., $X_2 = x_2$ have diameters between 0.4950 and 0.5050 in., and $X_3 = x_3, x_3 = n - x_1 - x_2$ have diameters greater than 0.5050 in.?
   (a) Find the marginal distribution function of X_2 and explain the result in words.
   (b) What are the mean and variance of X_2? of X_1?
   (c) Find the covariance of X_1 and X_2.

10. Suppose that 420 "true" dice are rolled simultaneously.
   (a) If X is a random variable denoting the total number of aces that turn up, find the values of the mean and standard deviation of X.
   (b) If Y is a random variable denoting the total number of dots that turn up, find the values of the mean and standard deviation of Y.

11. A process randomly generates digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with equal probabilities. If T is a random variable representing the sum of n digits taken from the process, find the mean and the variance of T.

12. A person plays 10 hands of bridge during the evening; suppose that T represents the total number of spades he obtains during the evening. Find the mean and variance of T.

13. A sample of size n is drawn from a lot of 10,000 articles known to contain 100 defectives. Using the Chebyshev's inequality, determine how large n should be in order for the probability to exceed 0.96 that the percentage of defectives in the sample will lie within the interval $(0.1, 1.9)$.

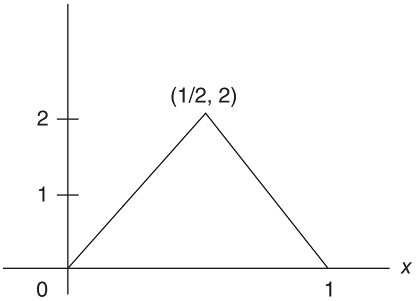
14. Using the Chebyshev's inequality, determine how many random digits should be generated in order for the probability to exceed 0.9000 that the mean of the random digits will lie within the interval $(3.5, 5.5)$.

15. Suppose that an insurance company has among all of its insurance policies, 50,000 policies for \$5000 for American men aged 51. The probability of an American male aged 51 dying within one year is (approximately) 0.01. Using the Chebyshev's inequality, decide for what value of k the probability exceeds 0.99 that the total death claims from the beneficiaries of this group (for the one year) will fall in the interval (\$2,500,000 $- k$, \$2,500,000 $+ k$).

16. If 2500 coins in a sack are poured out on a table, find with the use of the Chebyshev's inequality the value of k for which the probability that the number of heads will lie in the interval $(1250 - k, 1250 + k)$ exceeds 0.96.

17. One cigarette from each of four brands A, B, C, D is partially smoked by a blindfolded person. As soon as he takes a few puffs on a cigarette, he states the letter of the brand to which he considers it to belong. (Of course, he can use each letter only once.) Let X be the random variable denoting the number of cigarettes correctly identified. If the identification is done at random (i.e., he is equally likely to assign any letter to any cigarette), write down the probability distribution of X in table form. Find the mean and variance of X.

18. A point is taken at random from the interval $(0, 1)$, all points being equally likely. A second point is then taken in the same way. Let X be the coordinate of the point halfway between these points. X is a continuous chance quantity with a probability density function having an inverted V graph as shown below:



Write down the formula for $f(x)$. Find the mean and variance of X. Find the formula for $F(x)$ and graph $F(x)$.

19. By using the moment-generating function of a random variable X having the binomial distribution with parameter p, show that the mean and the variance of X are np and $np(1 - p)$, respectively.

20. Suppose that $X_1, \ldots, X_n$ is a sample from a distribution whose mean is μ, variance is σ^2, and whose moment-generating function exists. Show, using the method of moment-generating functions, that the mean and variance of the sample sum T are $n\mu$ and $n\sigma^2$, respectively. Also use the method of moment-generating functions to show that the mean and variance of the sample mean $\bar{X}$ are μ and σ^2/n, respectively.

21. If (X_1, X_2) is a pair of random variables such that

$$p_2(x_2) = \frac{\mu^{x_2} e^{-\mu}}{x_2!}, \quad x_2 = 0, 1, 2, \ldots$$

and

$$p(x_1|x_2) = \binom{x_2}{x_1} p^{x_1}(1-p)^{x_2-x_1}, \quad x_1 = 0, 1, 2, \ldots, x_2$$

show that $p_1(x_1)$ is a Poisson distribution.

22. Let X_1 be a number taken at random on the interval $(0, 1)$, and suppose that $X_1 = x_1$ is the observed value of X_1. Let X_2 be a number taken at random on the interval $(x_1, 1)$. Show that the distribution of X_2 has p.d.f.

$$f_2(x_2) = \begin{cases} -\ln(1-x_2), & \text{if} \quad 0 < x_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

23. Let $F(x, y)$ be the c.d.f. of random variables (X, Y). Show that

$$P(x_1 < X < x_2, y_1 < Y < y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \ge 0$$

24. Suppose that the random variables (X_1, X_2) are distributed as *bivariate normal* with parameters 0, 0, 1, 1, and ρ. Show that the random variables $X = X_1 + X_2$ and $Y = X_1 - X_2$ are independent.

25. Suppose that the random variables (X, Y) are *bivariate normal* with joint p.d.f.

$$f(x, y) = c\, exp\left\{-\frac{25}{18}\left(x^2 + y^2 + \frac{4}{5}x - \frac{14}{5}y - \frac{8}{5}xy + \frac{17}{5}\right)\right\}$$

   (a) Find the parameters of the bivariate normal
   (b) Find the value of c
   (c) Find the marginal p.d.f.'s of both X and Y.

26. Referring to Problem 25,
   (a) Find the conditional p.d.f. of the random variable Y, given $X = x$.
   (b) Find the conditional p.d.f. of the random variable X, given $Y = y$.
   (c) Find the conditional mean and variance of the random variable Y, given $X = x$.
   (d) Find the conditional mean and variance of the random variable X, given $Y = y$.

Chapter 7

SAMPLING DISTRIBUTIONS

The focus of this chapter is to discuss sampling distributions of functions of random variables.

Topics Covered

- Basic concepts of sampling from both an infinite and finite population
- Distributions of the sample average and sample proportion
- A fundamental theorem of probability, known as the "Central Limit Theorem"
- Distributions related to the normal distribution, namely chi-square, Student t, and Snedecor's F distributions, which are very useful in applied statistics
- Distributions of various order statistics and their applications
- Use of different statistical packages, namely MINITAB, R, and JMP

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the basic concepts of sampling distributions.
- Understand the Central Limit Theorem and when to apply it.
- Understand the details of important sampling distributions, namely χ^2-square, Student-t, and Snedecor's F-distributions and use them to make conclusions about problems that arise in applied statistics.
- Develop the distributions of various order statistics.
- Make use of MINITAB, R, and JMP in areas of applied statistics.

7.1 RANDOM SAMPLING

In this chapter, we study topics that build a foundation for what is called *inferential statistics*. In inferential statistics, we use the information contained in a random sample

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

to make predictions about the population from which the sample has been drawn. In Chapter 2, we had a very brief discussion about various designs of random sampling. However, we will now discuss the design for *simple random sampling* (or simply called *random sampling*) in detail.

7.1.1 Random Sampling from an Infinite Population

In this section, we discuss the special case that random variables $X_1, \ldots, X_n$ are mutually independent, and all have identical probability functions $p(x)$ (or identical probability density functions [p.d.f.'s] $f(x)$ in the case $X_1, \ldots, X_n$ are continuous random variables). This set of n random variables $(X_1, \ldots, X_n)$ is said to be a *random sample* from $p(x)$ (or from $f(x)$), and $X_1, \ldots, X_n$ are called *elements of the sample*, or observations. Thus, the sample elements $X_1, \ldots, X_n$ have equal means, which we will denote by μ, and equal variances, which we denote by σ^2. Sometimes we say that $(X_1, \ldots, X_n)$ is a random sample from a *population* having probability function (p.f.) $p(x)$, or p.d.f. $f(x)$, and μ and σ^2 are called the *population mean* and *population variance*, respectively. We often say that $X_1, \ldots, X_n$ are n independent observations on a random variable X, where $E(X) = \mu$ and $Var(X) = \sigma^2$.

Now suppose we consider the sample sum T, defined by

$$T = X_1 + \cdots + X_n \tag{7.1.1}$$

The reader may verify that (see Theorem 6.2.3) the mean and variance of T are given by

$$E(T) = n\mu \tag{7.1.2}$$

and

$$Var(T) = n\sigma^2 \tag{7.1.3}$$

respectively. Since $X_1, \ldots, X_n$ are n independent observations on a random variable X, the quantity T is also a random variable, and equations (7.1.2) and (7.1.3) give some important properties of T.

Now, the reader must bear in mind that T is the sum of n independent measurements $X_i, i = 1, \ldots, n$ and not equal to the variable $U = nX$. While it is true that $E(T) = E(U) = n\mu$, the variance of T is $n\sigma^2 \neq n^2\sigma^2 = Var(U)$. This occurs because U is found by observing a single X, and multiplying it by n. Recall that, T, however, is found by observing n "X's", which we are assuming to be independent observations, $X_1, \ldots, X_n$, and then summing the results, so that we have

$$T = X_1 + \cdots + X_n$$

Continuing, suppose that we still assume that $(X_1, \ldots, X_n)$ are n independent observations on X. We let $\bar{X}$ be the sample average, defined by

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \tag{7.1.4}$$

and it can be seen that the *mean* and *variance* of $\bar{X}$ are given by [see equations (6.2.17) and (6.2.19)]

$$E(\bar{X}) = \frac{1}{n}E(T) = \mu \tag{7.1.5}$$

$$Var(\bar{X}) = \frac{1}{n^2}\sigma_T^2 = \frac{\sigma^2}{n} \tag{7.1.6}$$

since $\sigma_T^2 = n\sigma^2$. Equations (7.1.5) and (7.1.6) state particularly important properties of $\bar{X}$. The fact that $E(\bar{X}) = \mu$ simply means the distribution of $\bar{X}$ has the same mean as the population from which the sample has been drawn. The fact that $Var(\bar{X})$ is inversely proportional to n shows that as n increases, the distribution of $\bar{X}$ becomes more highly concentrated about its mean μ.

Example 7.1.1 (Mean and variance of sum of n observations) *Suppose that* $(X_1, \ldots, X_n)$ *are observations on weights of n plastic items produced by a molding machine. If the mean and variance of the weights of an indefinitely large population of items molded by this particular machine using this kind of plastic are* μ *and* σ^2*, then the mean and variance of* $T = X_1 + \cdots + X_n$*, the total weight of the sample of n elements* $X_1, \ldots, X_n$*, are* $n\mu$ *and* $n\sigma^2$*, respectively.*

Example 7.1.2 (Mean and variance of sample mean) *Suppose that* $(X_1, \ldots, X_n)$ *are random variables representing errors made independently while measuring the length of a bar* n *times when its "true" length is known. If it is assumed that* $(X_1, \ldots, X_n)$ *are independent random variables, each with mean* 0 *and variance* σ^2*, then the mean and variance of the average error* $\bar{X}$ *are* 0 *and* σ^2/n*, respectively.*

Now by using Chebyshev's Inequality (5.3.1), we can make a stronger statement about the concentration of probability in the distribution of $\bar{X}$ around its mean μ than by merely saying that the variance of $\bar{X}$ is σ^2/n and, thus, inversely proportional to n. More precisely, for any given arbitrarily small number $\epsilon > 0$, we will consider

$$P(|\bar{X} - \mu| > \epsilon) \tag{7.1.7}$$

that is, the probability that $\bar{X}$ will fall *outside* the interval

$$[\mu - \epsilon, \mu + \epsilon]$$

We rewrite (7.1.7) as

$$P(|\bar{X} - \mu| > \epsilon) = P\left(|\bar{X} - \mu| > \left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\frac{\sigma}{\sqrt{n}}\right) \tag{7.1.7a}$$

Applying the Chebyshev's Inequality and noting that $\epsilon\sqrt{n}/\sigma$ in (7.1.7a) plays the role of k in (5.3.1) and $\sigma/\sqrt{n}$ is the standard deviation of $\bar{X}$ and plays the role of σ in (5.3.1), so that applying (5.3.1) we have the following result.

$$P(|\bar{X} - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n} \tag{7.1.8}$$

This states that, for an arbitrarily small positive number ϵ, the sample size n can be chosen sufficiently large to make the probability as small as we please so that the average $\bar{X}$ will not differ from the population mean μ by more than ϵ. Statement (7.1.8) is sometimes called the *weak law of large numbers.*

7.1.2 Random Sampling from a Finite Population

In Section 7.1.1, we discussed random sampling from *an infinite population* having the probability distribution with p.f. $p(x)$ or a p.d.f. $f(x)$. Now suppose that the population being sampled has only a finite number of objects say $O_1, O_2, \ldots, O_N$, such that the x-values of these objects are $X_1, X_2, \ldots, X_N$, respectively. We now define the mean μ and the variance σ^2 of this population as follows.

$$\mu = \frac{1}{N}(X_1 + X_2 + \cdots + X_N) \tag{7.1.9}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{7.1.10}$$

If we consider a sample $X_1, X_2, \ldots, X_n$ of size n drawn from the finite population of N objects, without replacement, there are $\binom{N}{n}$ possible samples that could be drawn, each of which would have a certain sum T and an average $\bar{X}$.

If each of these samples is equally likely to be drawn, that is, if the sample is a random sample, then the *mean* and *variance* of all $\binom{N}{n}$ possible sample sums are given by

$$E(T) = n\mu \tag{7.1.11}$$

$$Var(T) = \left(\frac{N-n}{N-1}\right) n\sigma^2 \tag{7.1.12}$$

The proofs of equations (7.1.11) and (7.1.12) are not given here but are available on the book website: www.wiley.com/college/gupta/statistics2e. From equations (7.1.11) and (7.1.12), it follows that

$$E(\bar{X}) = \mu \tag{7.1.13}$$

$$Var(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \tag{7.1.14}$$

Further by letting $N \to \infty$ it follows from equation (7.1.14) that $\lim_{N\to\infty} Var(\bar{X}) = \frac{\sigma^2}{n}$.

Example 7.1.3 (Mean and variance of sum of n random numbers) *If N objects are numbered $1, 2, \ldots, N$, respectively, and if a sample of n objects is drawn at random from this population of N objects, what is the mean and variance of the sum T of numbers drawn in the sample?*

In this example, the X values $X_1, X_2, \ldots, X_N$ of the objects may be taken as $1, 2, \ldots, N$ respectively. Thus, the mean and the variance of the population are given by

$$\begin{aligned}\mu &= (1 + 2 + \cdots + N)/N = (N + 1)/2\\ \sigma^2 &= \frac{1}{N}(1^2 + 2^2 + \cdots + N^2 - N\mu^2)\\ &= \frac{1}{N}\left[\frac{N(N+1)(2N+2)}{6} - N\left(\frac{N+1}{2}\right)^2\right]\end{aligned}$$

that is

$$\mu = \frac{1}{2}(N + 1), \quad \sigma^2 = \frac{1}{12}(N^2 - 1)$$

Hence, for a sample of n objects, T is such that

$$E(T) = \frac{n}{2}(N + 1) \text{ and } Var(T) = \frac{n}{12}(N^2 - 1)$$

PRACTICE PROBLEMS FOR SECTION 7.1

1. Define the appropriate population from which the following samples have been drawn:
   (a) Fifty employees from a manufacturing company are asked if they would like the company to have some training program for all employees.
   (b) A quality control engineer of a semiconductor company randomly selects 10 chips from a batch to examine their quality.
   (c) One hundred voters from a large metropolitan area are asked for their opinion about the location of the proposed airport.
2. The monthly entertainment expenses to the nearest dollar of 10 college students randomly selected from a university with 10,000 students are as follows:

48	46	33	40	29	38	37	37	40	48

   Determine the mean and standard deviation of these data.
3. Refer to Problem 2. Let T denote the total expenses for entertainment of all the students. Estimate the mean and variance of T.
4. A manufacturing company has developed a new device for the army, obtaining a defense contract to supply 25,000 pieces of this device to the army. In order to meet the contractual obligations, the department of human resources wants to estimate the number of workers that the company would need to hire. This can be accomplished by estimating the number of worker hours needed to manufacture 25,000

devices. The following data give the number of hours spent by randomly selected workers to manufacture 15 pieces of such a device.

8.90	8.25	6.50	7.25	8.00	7.80	9.90	8.30
9.30	6.95	8.25	7.30	8.55	7.55	7.70	

Estimate the total worker hours needed to manufacture 25,000 devices.

5. The following data give the scores on a midterm test of 20 randomly selected students.

29	26	40	27	35	39	37	40	37	34
36	28	26	33	37	25	27	33	26	29

Find the mean and standard deviation for these data.

6. Refer to Problem 5. Suppose that the instructor of the class decided to give 10 extra points to every student. Find the mean and variance of the new data and comment on your result.

7.2 THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Often we encounter processes of interest whose mean μ and variance σ^2 are not known. Hence, one of the problems in statistical inference is to estimate them. For example, we use the sample average $\bar{X}$ and the sample variance S^2 to estimate the population mean μ and population variance σ^2, respectively. However, it is important to note that the values of the statistics $\bar{X}$ and S^2 vary from sample to sample. So a question that then arises, is what guarantees are there that $\bar{X}$ and S^2 will give good estimates for μ and σ^2? To answer this question, it is important that we know the probability distributions of these statistics. The goal of the remainder of this chapter is to study the probability distributions of $\bar{X}$, S^2, and other related distributions. The probability distributions of various statistics are called their *sampling distributions.*

In this section, we consider the sampling distribution of the sample average $\bar{X}$ when (i) the sampled population is normal and (ii) the sampled population is nonnormal.

7.2.1 Normal Sampled Population

If the random sample $X_1, X_2, \ldots, X_n$ is taken from a normal population with mean μ and variance σ^2, then from Theorem 6.4.3, it follows that for $n \geq 1$, $\bar{X}$ is normally distributed with mean μ and variance σ^2/n, or standard deviation $\sigma/\sqrt{n}$, which is also called the *standard error.*

7.2.2 Nonnormal Sampled Population

Now, we discuss the distribution of $\bar{X}$ when we are sampling from a population that is nonnormal, either finite or infinite. In this case, the approximate distribution of $\bar{X}$ is given by a very important theorem of probability theory, known as the *central limit theorem.* We discuss this theorem next.

7.2.3 The Central Limit Theorem

Theorem 7.2.1 (Central Limit Theorem) *If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with mean μ and variance σ^2 (both finite), then the limiting form of the distribution of $Z_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is that of the standard normal, that is, normal with mean* 0 *and variance* 1.

The proof of the Central Limit Theorem is not given here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Figure 7.2.1 gives the distribution of the exponential population with mean 1 from which samples of various sizes ($n = 5, 10, 20, 230, 50,$ and 100) have been taken. Figure 7.2.2 gives the normal probability plots (see Section 5.8) of $\bar{X}$ for the selected sample sizes.

These normal probability plots clearly show that as the sample size increases, the sampling distribution of $\bar{X}$ approaches the normal distribution, even though the distribution of the population from which the samples have been taken is highly skewed.

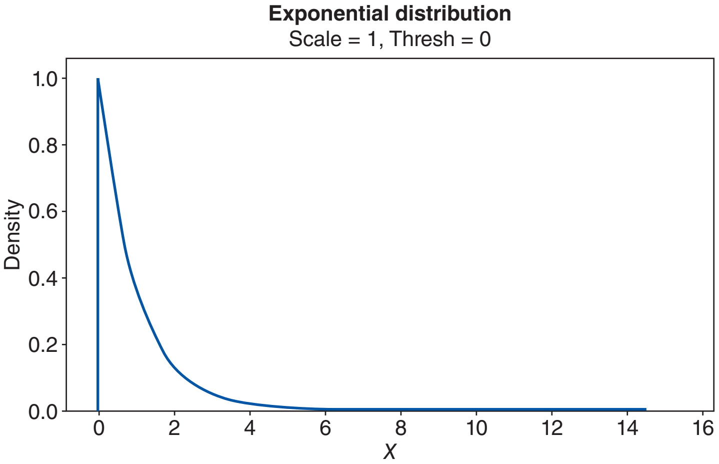


Figure 7.2.1 Exponential probability distribution of X with mean 1.

The normal approximation for the sampling distribution of $\bar{X}$ is generally considered acceptable if $n \geq 30$. For $n < 30$, the approximation is accurate only if the distribution of the population from which the samples are being taken is close to being normally distributed. If the population is normal, then the sampling distribution of $\bar{X}$ is exactly normal, regardless of the sample size; the sample size could even be as small as 2.

Example 7.2.1 (Applying the Central Limit Theorem) *Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli population with parameter p, and suppose that $Y = X_1 + \cdots + X_n$. Find the sampling distribution of Y as $n \to \infty$.*

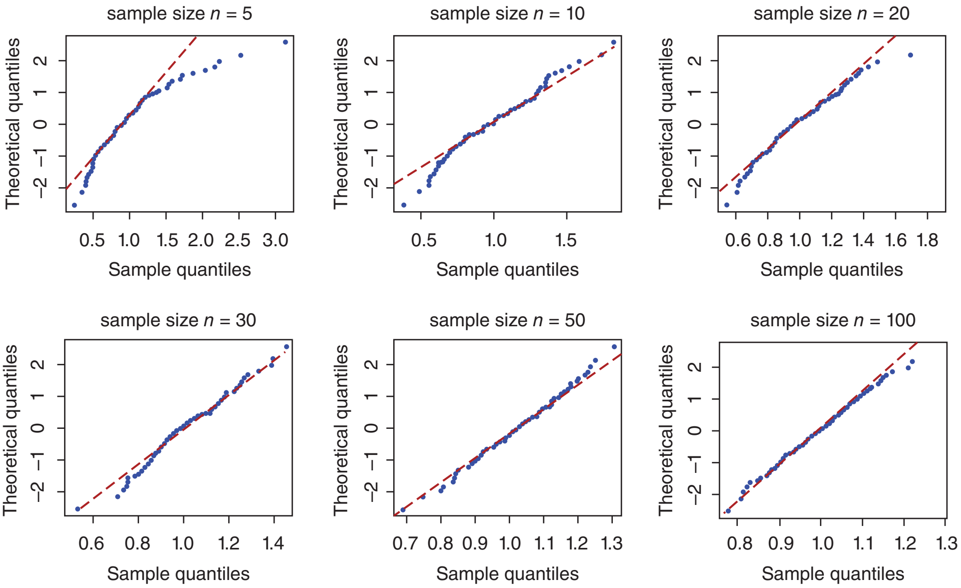


Figure 7.2.2 Normal probability plots of $\bar{X}$ for selected sample sizes $n = 5, 10, 20, 30, 50,$ and 100, when sampling from the exponential distribution with mean 1.

Solution: From Chapter 4, it follows that Y is a random variable having the binomial distribution with mean np and variance $np(1-p)$. Now, we also have that Y is the sum of n independent and identically distributed random variables, each having mean p and variance $p(1-p)$. Hence, by the central limit theorem, $(Y - np)/\sqrt{np(1-p)}$ is a random variable whose limiting distribution, as $n \to \infty$, is the standard normal distribution, so that the random variable Y is, for large n, approximately distributed as normal with mean np and variance $np(1-p)$. As we noted in Section 5.7, this normal approximation for the binomial is quite accurate when n and p are such that both the inequalities $np > 5$ and variance $n(1-p) > 5$ hold.

Note that if the sampling is done *without replacement* from a finite population with mean μ and variance σ^2, then the sampling distribution of $\bar{X}$ is still approximately normally distributed with mean μ and variance $\dfrac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$; the factor $\frac{N-n}{N-1}$ is called the *finite population correction factor*. Further note that this correction factor is ignored if either the sampling is done with replacement or the sample size relative to the population size is small $(n < 0.05N)$.

Example 7.2.2 (Applying the central limit theorem) *The mean weight of a food entrée is $\mu = 190$ g with a standard deviation of 14 g. If a random sample of 49 entrées is selected, then find*

(a) The probability that the sample average weight will fall between 186 and 194g.
(b) The probability that the sample average will be greater than 192 g.

Solution: (a) Let $\bar{X}$ be the sample mean. Since the sample size $n = 49$ is large, from the Central Limit Theorem, we know that $\bar{X}$ is approximately normal with mean $\mu = 190$ g and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 14/\sqrt{49} = 2$ g. Thus, we have

$$P(186 \leq \bar{X} \leq 194) = P\left(\frac{186-190}{2} \leq \frac{\bar{X}-190}{2} \leq \frac{194-190}{2}\right)$$

which to good approximation ($n = 49$ is large) is equal to

$$P(-2 \leq Z \leq 2) = P(Z \leq 2.0) - P(Z \leq -2.0) = 0.9772 - 0.0228 = 0.9544$$

(b) By the same argument as in part (a), we have, approximately, that

$$P(\bar{X} \geq 192) = P\left(\frac{\bar{X}-190}{2} \geq \frac{192-190}{2}\right) = P(Z \geq 1) = 0.1587$$

Example 7.2.3 (Example 7.2.2 continued) *Repeat Example 7.2.2 when the sample size is increased from 49 to 64 entrées.*

Solution: (a) In this example, $\bar{X}$ will be approximately normal with mean $\mu = 190$ g and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 14/\sqrt{64} = 1.75$ g.

Thus, we have

$$\begin{aligned} P(186 \leq \bar{X} \leq 194) &= P\left(\frac{186-190}{1.75} \leq \frac{\bar{X}-190}{1.75} \leq \frac{194-190}{1.75}\right) \\ &= P(-2.28 \leq Z \leq 2.28) = P(Z \leq 2.28) - P(Z \leq -2.28) = 0.9774 \end{aligned}$$

(b) In this case, we have

$$P(\bar{X} \geq 192) = P\left(\frac{\bar{X}-190}{1.75} \geq \frac{192-190}{1.75}\right) = P(Z \geq 1.14) = 0.1271$$

From Examples 7.2.2 and 7.2.3, we see that as the sample size increases from 49 to 64, the probability that the sample mean falls within 4 units of the population mean increases, while the probability that the sample mean falls beyond two units from the population mean decreases. These examples emphasize that as the sample size increases, the variance of the sample mean decreases, and therefore, the sample means are more concentrated about the population mean. This fact makes $\bar{X}$ a good estimator for the population mean μ. We will study this phenomenon in more detail in Chapter 8.

Example 7.2.4 (Applying the central limit theorem to approximate probabilities for a sample mean) *A random sample of 36 reinforcement rods is taken from a production plant that produces these rods with a mean length of $\mu = 80$ cm and a standard deviation of 0.6 cm. Find the approximate probability that the sample mean of the 36 rods falls between 79.85 and 80.15 cm.*

Solution: Let $\bar{X}$ be the sample mean. We are then interested in finding the probability of $P(79.85 \leq \bar{X} \leq 80.15)$.

Since the sample size is large, by the central limit theorem, we know that $\bar{X}$ is approximately normally distributed with mean 80 cm and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.6/\sqrt{36} = 0.1$ cm. Thus we have

$$P(79.85 \leq \bar{X} \leq 80.15) = P\left(\frac{79.85 - 80}{0.1} \leq \frac{\bar{X} - 80}{0.1} \leq \frac{80.15 - 80}{0.1}\right)$$

which to good approximation is equal to

$$P\left(-\frac{0.15}{0.1} \leq Z \leq \frac{0.15}{0.1}\right) = P(-1.5 \leq Z \leq 1.5) = P(Z \leq 1.5) - P(Z \leq -1.5) = 0.8664$$

Example 7.2.5 (Hourly wages of workers in a semiconductor industry) *Suppose that the mean hourly wage of all employees in a large semiconductor manufacturing facility is \$50.00 with a standard deviation of \$10.00. Let $\bar{X}$ be the mean hourly wages of certain employees selected randomly from all the employees of this manufacturing facility. Find the approximate probability that the mean hourly wages $\bar{X}$ falls between \$48.00 and \$52.00 when the number of selected employees is (a) 64, (b) 100.*

Solution: (a) Since the sample size of 64 is large, and by the central limit theorem, we know that $\bar{X}$ is approximately normally distributed with mean $\mu_{\bar{x}} = \mu = 50$, and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{64} = 1.25$. We are interested in finding the probability $P(48 \leq \bar{X} \leq 52)$, which, approximately, is given by

$$\begin{aligned} P(48 \leq \bar{X} \leq 52) &= P\left(\frac{48 - 50}{1.25} \leq \frac{\bar{X} - 50}{1.25} \leq \frac{52 - 50}{1.25}\right) \\ &= P(-1.6 \leq Z \leq 1.6) = P(Z \leq 1.6) - P(Z \leq -1.6) = 0.8904 \end{aligned}$$

(b) In this case, the sample size is 100, which is also large. By the same argument as used in part (a), we have

$$\mu_{\bar{x}} = \mu = 50 \quad \text{and} \quad \sigma_{\bar{x}} = 10/\sqrt{100} = 1$$

Thus, the desired probability is given by

$$\begin{aligned} P(48 \leq \bar{X} \leq 52) &= P\left(\frac{48 - 50}{1} \leq \frac{\bar{X} - 50}{1} \leq \frac{52 - 50}{1}\right) \\ &= P(-2 \leq Z \leq 2) = P(Z \leq 2.0) - P(Z \leq -2.0) = 0.9544 \end{aligned}$$

Example 7.2.6 (Example 7.2.5 continued) *Repeat Example 7.2.5 for the company that is not very large and whose total number of employees is only 500.*

Solution: In this case, the sample size is large, but the population is finite. Before applying the Central Limit Theorem, we must check whether $n < (0.05)N$. If this relation does not hold, we must use the finite population correction factor to calculate the standard deviation of the sample average.

(a) In this case, we have $n/N = 64/500 = 0.128$, so that the sample size $n = 64$ is greater than 5% of the population. Using the finite population correction factor, we have

$$\mu_{\bar{x}} = \mu = 50$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{10}{8} \times \sqrt{\frac{500-64}{500-1}} = 1.25\sqrt{\frac{436}{499}} = 1.168$$

Therefore, the desired probability is approximately

$$\begin{aligned} P(48 \le \bar{X} \le 52) &= P\left(\frac{48-50}{1.168} \le \frac{\bar{X}-50}{1.168} \le \frac{52-50}{1.168}\right) \\ &= P(-1.71 \le Z \le 1.71) = P(Z \le 1.71) - P(Z \le -1.71) = 0.9128 \end{aligned}$$

(b) Again, note that the sample size is greater than 5% of the population size. Thus, using the finite population correction factor, we have

$$\mu_{\bar{x}} = \mu = 50$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} = \frac{10}{10} \times \sqrt{\frac{500-100}{500-1}} = 0.895$$

Therefore, the desired probability is approximated by

$$\begin{aligned} P(48 \le \bar{X} \le 52) &= P\left(\frac{48-50}{0.895} \le \frac{\bar{X}-50}{0.895} \le \frac{52-50}{0.895}\right) \\ &= P(-2.23 \le Z \le 2.23) = P(Z \le 2.23) - P(Z \le -2.23) = 0.9742 \end{aligned}$$

Note that in this example, both probabilities are slightly greater than those found in Example 7.2.5. This is due to our using the finite population correction factor by which the standard deviation of $\bar{X}$ becomes smaller.

Example 7.2.7 (Sampling distribution of the estimator of the binomial parameter) *Let X be a random variable distributed as binomial with parameters n and p, where n is the number of trials, and p is the probability of success. Find the sampling distribution of the sample proportion $\hat{p}$ when (a) $n = 100, p = 0.25$ and (b) $n = 64, p = 0.5$.*

Solution: (a) We have $np = 100(0.25) = 25 > 5$, and $n(1-p) = 100(1-0.25) = 75 > 5$. By applying the central limit theorem, we see that $\hat{p}$ is approximately normally distributed with mean and variance given by

$$\mu_{\hat{p}} = p = 0.25, \quad \text{and} \quad \sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{(0.25)(0.75)}{100} = 0.001875$$

(b) By the same argument as in part (a), we have $np = 64(0.5) = 32 > 5$ and $n(1-p) = 64(1-0.5) = 32 > 5$

Again, we see that $\hat{p}$ is approximately normally distributed with mean and variance given by

$$\mu_{\hat{p}} = p = 0.5, \quad \text{and} \quad \sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{(0.5)(0.5)}{64} = 0.0039$$

PRACTICE PROBLEMS FOR SECTION 7.2

1. Suppose that random samples of size 36 are repeatedly drawn from a population with mean 28 and standard deviation 9. Describe the sampling distribution of $\bar{X}$.
2. Suppose that random samples (without replacement) of size 5 are repeatedly drawn from a finite population of size $N = 50$. Suppose that the mean and the standard deviation of the population are 18 and 5, respectively. Find the mean and the standard deviation of the sampling distribution of $\bar{X}$.
3. Suppose that random samples are drawn from an infinite population. How does the standard error of $\bar{X}$ change if the sample size is (i) increased from 36 to 64, (ii) increased from 100 to 400, (iii) increased from 81 to 324, (iv) increased from 256 to 576?
4. The weight of all cars traveling on interstate highway I-84 is normally distributed with a mean of 3000 lb and standard deviation of 100 lb. Let $\bar{X}$ be the mean weight of a random sample of 16 cars traveling on I-84. Calculate the probability that $\bar{X}$ falls between 2960 and 3040 lb.
5. Suppose that the amount of a weekly grocery bill of all households in a metropolitan area is distributed with a mean of \$140 and a standard deviation of \$35. Let $\bar{X}$ be the average amount of grocery bill of a random sample of 49 households selected from this metropolitan area. Find the probability that $\bar{X}$ will be (a) more than \$145, (b) less than \$140, (c) between \$132 and \$148.
6. Let X be a random variable distributed as binomial $B(n, p)$. State the approximate sampling distribution of the sample proportion $\hat{p}$ when (a) $n = 40, p = 0.4$; (b) $n = 50, p = 0.2$; (c) $n = 80, p = 0.1$.
7. In 1995, the median price of a PC was \$1200. Suppose that a random sample of 100 persons who bought their PCs during that year recorded the amount spent (by each of them) on his/her PC. State the approximate sampling distribution of $\hat{p}$, the proportion of persons who spent more than \$1200 on a PC. Find the probability that more than 60% of this group spent more than \$1200.
8. The amount of beverage dispensed by a bottling machine is normally distributed with mean of 12 oz and a standard deviation of 1 oz. A random sample of n bottles is selected, and a sample average $\bar{X}$ is calculated. Determine the following probabilities:
   (a) $P(|\bar{X} - 12| \leq 0.25)$ for sample sizes $n = 16, 25, 36, 49$, and 64.
   (b) Comment on the values of probabilities obtained in part (a).

7.3 SAMPLING FROM A NORMAL POPULATION

In this section, we consider various sampling distributions that arise when sampling from a normal population. These distributions are widely used in applied statistics.

7.3.1 The Chi-Square Distribution

We encounter the *chi-square distribution*, usually denoted as the χ^2-distribution, quite frequently in statistical applications. The χ^2-distribution occupies an important place in applied statistics. The χ^2-distribution moreover is related to the normal distribution, as discussed later in this section. We start with the following definition:

Definition 7.3.1 A random variable W is said to be distributed as a χ^2_n random variable if its p.d.f. is given by

$$f(w) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} w^{n/2-1} e^{-w/2}, & w \geq 0, n > 0 \\ 0, & \text{otherwise} \end{cases}$$

We sometimes say that W is a *chi-square random variable* with n degrees of freedom and denote this by $W \sim \chi^2_n$, which is read as W is distributed as the chi-square random variable with n degrees of freedom.

Theorem 7.3.1 *The moment-generating function of W, where $W \sim \chi^2_n$, is given by*

$$M_W(t) = M_{\chi^2_n}(t) = (1 - 2t)^{-n/2}$$

Proof: By definition, the moment-generating function of W is

$$M_W(t) = E(e^{tW}) = \int_0^\infty e^{tw} f(w) dw$$

Now $f(w)$ is defined in Definition 7.3.1, so we may write

$$M_W(t) = \int_0^\infty \frac{1}{2^{n/2}\Gamma(n/2)} w^{n/2-1} e^{-\frac{w}{2}(1-2t)} dw$$

Assume that $(1 - 2t) > 0$. Then, if we let $u = ((1 - 2t)/2)w$ in the above integral, we have, after using some elementary algebra that

$$M_W(t) = \frac{2^{n/2}\Gamma(n/2)}{2^{n/2}\Gamma(n/2)} (1 - 2t)^{-n/2}$$

that is,

$$M_W(t) = M_{\chi^2_n}(t) = (1 - 2t)^{-n/2} \tag{7.3.1}$$

□

Theorem 7.3.2 *Let $Z_1, \ldots, Z_n$ be a random sample from a standard normal distribution $N(0, 1)$. Let a new random variable* Y *be defined as follows:*

$$Y = Z_1^2 + \cdots + Z_n^2 \tag{7.3.2}$$

Then, the random variable Y is distributed as chi-square *with* n *degrees of freedom and is written as χ^2_n.*

Proof: Since the random variable Z_i is distributed as standard normal, the moment-generating function (m.g.f.) of Z_i^2 is given by

$$\begin{aligned} M_{Z_i^2}(t) = E(e^{tZ_i^2}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz_i^2} e^{-z_i^2/2} dz_i \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2t)z_i^2/2} dz_i \\ &= (1-2t)^{-1/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du \end{aligned}$$

where $u^2 = (1-2t)z_i^2$. We then have

$$M_{Z_i^2}(t) = (1-2t)^{-1/2}$$

This is the m.g.f. of a χ_1^2 random variable (see Theorem 7.3.1), so that $Z_i^2 \sim \chi_1^2$. Furthermore, since the random variables $Z_1, \ldots, Z_n$ are independent and identically distributed (see Chapter 6), the m.g.f. of the random variable Y of Theorem 7.3.2 is given by

$$M_Y(t) = \prod_{i=1}^{n} M_{Z_i^2}(t) = (1-2t)^{-n/2} \tag{7.3.3}$$

From Theorem 7.3.1, the m.g.f. of the random variable Y is the m.g.f. of χ_n^2 random variable, so that $Y \sim \chi_n^2$, which in turn implies that the p.d.f. of Y is

$$f(y) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, & y \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{7.3.4}$$

□

The chi-square distribution has only one parameter n, the degrees of freedom. The shape of various p.d.f. of χ_n^2 for $n = 4, 6, 8$, and 10 are shown in Figure 7.3.1.

Note that from the definition of the χ^2 p.d.f., the random variable χ^2 assumes only nonnegative values. Hence, the entire frequency distribution curve falls to the right of the origin and it is right-skewed. The *mean* and *variance* of the chi-square distribution are, respectively, equal to the degrees of freedom and twice the degrees of freedom. That is, for χ_n^2 we have that the mean and variance are as given below.

$$\mu_{\chi_n^2} = n \tag{7.3.5}$$

$$\sigma_{\chi_n^2} = 2n \tag{7.3.6}$$

We also state the following important fact as a corollary below.

Corollary 7.3.1 *Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, \sigma^2)$, and let $\bar{X}$ be the sample average. Then, the random variable $(\bar{X} - \mu)^2/(\sigma^2/n)$ is distributed as a chi-square random variable with 1 degrees of freedom.*

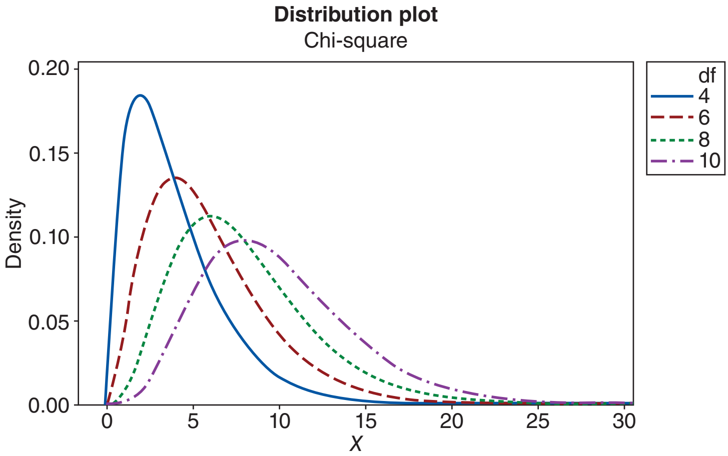


Figure 7.3.1 Graphs of chi-square density functions for different degrees of freedom.

The result stated in corollary (7.3.1) is quite evident because $\bar{X}$ is distributed as normal with mean μ and variance σ^2/n so that $(\bar{X}-\mu)/(\sigma/\sqrt{n}) \sim Z$; hence $[(\bar{X}-\mu)/(\sigma/\sqrt{n})]^2 \sim Z^2 \sim \chi_1^2$ (see Theorem 7.3.2). Also, we may state other facts related to chi-square variables below.

Corollary 7.3.2 *Let $X_1, \ldots, X_n$ be a random sample from the normal distribution $N(\mu, \sigma^2)$, then the random variable*

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{7.3.7}$$

is distributed as a chi-square with n degrees of freedom.

Definition 7.3.2 Suppose that $X_1, \ldots, X_n$ is a random sample from the normal distribution $N(\mu, 1)$. Then, the random variable $Y=\sum_{i=1}^{n} X_i^2$ is said to be distributed as a noncentral χ_n^2 with parameter $\delta = n\mu^2$. (δ is called the noncentrality parameter.)

Theorem 7.3.3 *If $\chi_{n_1}^2$ and $\chi_{n_2}^2$ are independent random variables having chi-square distributions with n_1 and n_2 degrees of freedom, respectively, then $\chi_{n_1}^2+\chi_{n_2}^2$ is a random variable having the chi-square distribution with n_1+n_2 degrees of freedom.*

We have $M_{\chi_{n_1}^2+\chi_{n_2}^2}(t) = M_{\chi_{n_1}^2}(t) \times M_{\chi_{n_2}^2}(t)$, so $M_{\chi_{n_1}^2+\chi_{n_2}^2}(t) = (1-2t)^{-n_1/2} \times (1-2t)^{-n_2/2} = (1-2t)^{-(n_1+n_2)/2}$, which is the moment-generating function of a $\chi_{n_1+n_2}^2$ random variable. Now invoking Theorem 6.4.1, we have that $\sum_{i=1}^{2}\chi_{n_i}^2 \sim \chi_{n_1+n_2}^2$.

The values of $\chi^2_{n,\alpha}$ such that $P(\chi^2_n \geq \chi^2_{n,\alpha}) = \alpha$ for various values of n and α are given in Table A.6. Note that $\chi^2_{n,\alpha}$ is sometimes called the upper 100 α percent point of the χ^2_n distribution. For example, if the random variable χ^2_n is distributed with $n = 18$, and if $\alpha = 0.05$, then from Table A.6, we find that

$$\chi^2_{18,0.05} = 28.8693$$

that is,

$$P(\chi^2_{18} > 28.8693) = 0.05$$

Note that to find the value of the random variable χ^2_n such that the lower tail area is α, we would use Table A.6 to find the value of $\chi^2_{n,1-\alpha}$ for selected values of α and n. We then would have that $P(\chi^2_n > \chi^2_{n,1-\alpha}) = 1 - \alpha$, so $P(\chi^2_n \leq \chi^2_{n,1-\alpha}) = \alpha$.

Example 7.3.1 (Finding tail probabilities of χ^2-distribution) *Suppose that we are dealing with χ^2_{20} and wish to find the value of* a *and* b *such that $P(\chi^2_{20} \geq a) = 0.05$ and $P(\chi^2_{20} \leq b) = 0.025$.*

Solution: We are given the area under the upper tail. We can find the value of a directly from the Table A.6 with $n = 20$ and $\alpha = 0.05$, that is,

$$a = \chi^2_{20,0.05} = 31.410$$

For the lower tail, we are given that the area under the lower tail of the p.d.f. of the χ^2_{20} distribution is 0.025. We have

$$0.025 = P(\chi^2_{20} \leq b) = 1 - P(\chi^2_{20} > b)$$

so that

$$P(\chi^2_{20} > b) = 0.975$$

Hence, from Table A.6, we find that

$$b = \chi^2_{20,0.975} = 9.591$$

In applications, most often we are interested in finding the distribution of the sample variance S^2 when a random sample is taken from a normal population. However, before discussing the distribution of S^2, we state below another important result in Theorem 7.3.4. The proof of this theorem is, however, beyond the scope of this book.

Theorem 7.3.4 *Let $X_1, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Let the random variables $\bar{X}$ and S^2 be the sample average and sample variance, respectively. Then, the random variables $\bar{X}$ and S^2 are independent.*

We are now able to state the result about the distribution of the sample variance S^2.

Theorem 7.3.5 *Let $X_1, \dots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Consider*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{7.3.8}$$

the sample variance. Then, the random variable

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{7.3.9}$$

is distributed as χ^2 with $(n-1)$ degrees of freedom.

Example 7.3.2 (Finding the middle 95% interval for σ^2) *Suppose that a tea-packaging machine is calibrated such that the amount of tea it discharges is normally distributed with mean $\mu = 16$ oz and standard deviation $\sigma = 1.0$ oz. Suppose that we randomly select 21 packages and weigh the amount of tea in each package. If the sample variance of these 21 weights is denoted by S^2, then it may be of interest to find the values of c_1 and c_2 such that $P(c_1 \leq S^2 \leq c_2) = 0.95$, where $P(S^2 < c_1) = P(S^2 > c_2) = 0.025$, so (c_1, c_2) contains the middle or central part of the distribution of S^2.*

The solution to this problem would enable us to calibrate the machine such that the value of the sample variance would be expected to fall between certain values with a very high probability.

Solution: From Theorem 7.3.5, we have the following:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{20}$$

Now, we wish to find c_1 and c_2 as previously mentioned so that $P(c_1 \leq S^2 \leq c_2) = 0.95$, or

$$P\left(\frac{n-1}{\sigma^2}c_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{n-1}{\sigma^2}c_2\right) = 0.95$$

or

$$P\left(\frac{n-1}{\sigma^2}c_1 \leq \chi^2_{20} \leq \frac{n-1}{\sigma^2}c_2\right) = 0.95$$

since $n = 21$. But here $\sigma = 1$, so we now have $(n - 1 = 21 - 1 = 20)$

$$P(20c_1 \leq \chi^2_{20} \leq 20c_2) = 0.95$$

Then, by assigning probability 0.025 under each tail, or equivalently selecting the *middle* 0.95 of the χ^2_{20} distribution, we have from Table A.6 that $P(\chi^2_{20} \leq \chi^2_{20,0.975} = 9.951) = 0.025$ and $P(\chi^2_{20} \geq \chi^2_{20,0.025} = 34.170) = 0.025$, so that

$$P(9.591 \leq \chi^2_{20} \leq 34.170) = 0.95$$

and we have

$$20c_1 = 9.591 \quad \text{and} \quad 20c_2 = 34.170$$

or

$$c_1 = 0.4795 \quad \text{and} \quad c_2 = 1.7085$$

Thus,

$$P(0.4795 \leq S^2 \leq 1.7085) = 0.95$$

Note that from equation (7.3.9) of Theorem 7.3.5 and using equation (7.3.5), we can easily prove another important result given below.

$$E(S^2) = \sigma^2 \qquad (7.3.10)$$

Example 7.3.3 (Using MINITAB and R to find chi-square probabilities) *Using MINITAB and R determine the following:*

(a) Values of $\chi^2_{20,1-\alpha}$ *for* $\alpha = 0.01, 0.025, 0.05.$
(b) Values of $\chi^2_{20,\alpha}$ *for* $\alpha = 0.01, 0.025, 0.05.$

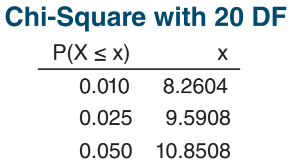
Solution:

MINITAB

(a) Recall that $P(\chi^2_{20} \geq \chi^2_{20,1-\alpha}) = 1 - \alpha$ so that $P(\chi^2_{20} \leq \chi^2_{20,1-\alpha}) = 1 - (1 - \alpha) = \alpha$. That is, $\chi^2_{20,1-\alpha}$ is the lower $100\alpha\%$ point of the χ^2_{20} distribution. In order to determine the value of $\chi^2_{20,1-\alpha}$ for $\alpha = 0.01, 0.025, 0.05$, we proceed as follows (*note that MINITAB determines the areas under the lower tail*):

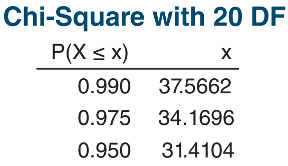
1. Enter the values 0.01, 0.025, 0.05 in column C1.
2. From the Menu bar select **Calc** > **Probability Distribution** > **Chi-square**.
3. In the dialog box that appears, click the circle next to **Inverse probability**.
4. Complete the boxes next to **Degrees of freedom** and **Input Column**, and click **OK**.

   The values of $\chi^2_{20,1-\alpha}$ will appear in the Session window shown below:



```
Chi-Square with 20 DF
P(X ≤ x)        x
   0.010   8.2604
   0.025   9.5908
   0.050  10.8508
```

(b) We first recall that $P(\chi^2_{20} \geq \chi^2_{20,\alpha}) = \alpha$, so $P(\chi^2_{20} \leq \chi^2_{20,\alpha}) = 1 - \alpha$. As previously mentioned, MINITAB determines values of $\chi^2_{20,\alpha}$ that are such that the area under the lower tail of the χ^2_{20} distribution is $1-\alpha$. Hence, we first enter the values of $1-\alpha$, that is, 0.99, 0.975, and 0.95, and then proceed in the same manner as in part (a). We find for this part that the values of $\chi^2_{20,\alpha}$ are:



Chi-Square with 20 DF

P(X ≤ x)	x
0.990	37.5662
0.975	34.1696
0.950	31.4104

USING R

R has a built-in chi-square distribution function ‘qchisq(p, df, ncp = 0, lower.tail = TRUE)’, where p is the probability, df is the degrees of freedom, ncp is the noncentrality parameter, and lower.tail = TRUE gives the quantile corresponding to the left tail chi-square probability p. So, referring to Example 7.3.3, in part (a) we use lower.tail = TRUE and in part (b) lower.tail = FALSE as we need lower and upper tail quantiles, respectively. Run following R code in the R Console window to obtain required probabilities as shown below.

```
prob = c(0.01, 0.025, 0.05)
qchisq(prob, df=20, ncp = 0, lower.tail = TRUE)
#R output
[1] 8.260398 9.590777 10.850811

qchisq(prob, df=20, ncp = 0, lower.tail = FALSE)
#R output
[1] 37.56623 34.16961 31.41043
```

7.3.2 The Student *t*-Distribution

One of the most important distributions in statistics, called the Student *t*-distribution, arises in problems involving small samples from the normal distribution. We have the following theorem:

Theorem 7.3.6 *If* X *and* Y *are independent random variables having the normal distribution $N(0,1)$ and the chi-square distribution with n degrees of freedom, respectively, then the random variable*

$$T = \frac{X}{\sqrt{Y/n}} \tag{7.3.11}$$

has the following probability density function:

$$f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty \leq t \leq \infty \tag{7.3.12}$$

The derivation of the probability distribution in equation (7.3.12) is not given here but is available on the book website: www.wiley.com/college/gupta/statistics2e.

It can be shown that the mean and the variance of T are given by

$$\mu = E(T) = 0 \quad \text{and} \quad \sigma^2 = Var(T) = \frac{n}{n-2} \tag{7.3.13}$$

the variance being finite only if $n > 2$.

Definition 7.3.3 The distribution having the probability density function (7.3.12) is called the Student t-distribution with n degrees of freedom. A random variable T having (7.3.12) as its probability density function is called a Student t-variable with n degrees of freedom and is often denoted by t_n

The p.d.f. of t_n is symmetric about zero. Its graph looks something like that of the p.d.f. of the normal distribution $N(0,1)$ shown in Figure 7.3.3. Theoretically, both distributions extend between $-\infty$ and ∞, but in practice, for the normal distribution almost all the probability falls between -3 and 3, whereas for the t-distribution, probabilities depend on the degrees of freedom. For example, for 15 degrees of freedom, almost all the distribution falls between -4 and 4. Thus, the t-distribution is slightly flatter than the normal distribution. However, it can be proved that as $n \to \infty$, the p.d.f. (7.3.12) tends to the p.d.f. of the $N(0,1)$ variable as its limit. Values of $t_{n,\alpha}$ for which

$$P(t_n > t_{n,\alpha}) = \alpha$$

are tabulated in Table A.5 for various values of n and α. Because of the symmetry around zero, we have that $t_{n,\alpha} = -t_{n,1-\alpha}$. That is, $P(t_n > t_{n,\alpha}) = P(t_n < -t_{n,\alpha}) = \alpha$ (see Figure 7.3.2). Note in Figures 7.3.2 and 7.3.3 that the tail areas of the t-distribution with the same probabilities are located farther from the origin (mean) than in the normal distribution, supporting our assertion that t-distribution is flatter than the normal distribution.

Example 7.3.4 (Using MINITAB and R to determine probabilities of the t-distribution) *Using MINITAB and R, determine the values of* $t_{25,1-\alpha}$ *such that:*

1. $P(t_{25} < t_{25,1-\alpha}) = \alpha$ *for* $\alpha = 0.01, 0.025, 0.05$ *(Recall that the t p.d.f. is symmetric around zero).*
2. *Values of* $t_{25,\alpha}$ *for which* $P(t_{25} > t_{25,\alpha}) = \alpha$ *for* $\alpha = 0.01, 0.025, 0.05$

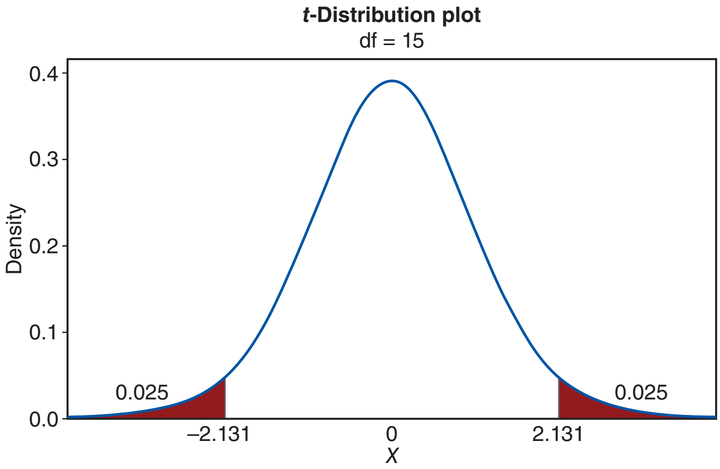


Figure 7.3.2 Probability density function of the t-distribution with $n = 15$ degrees of freedom showing 0.025 areas under each tail.

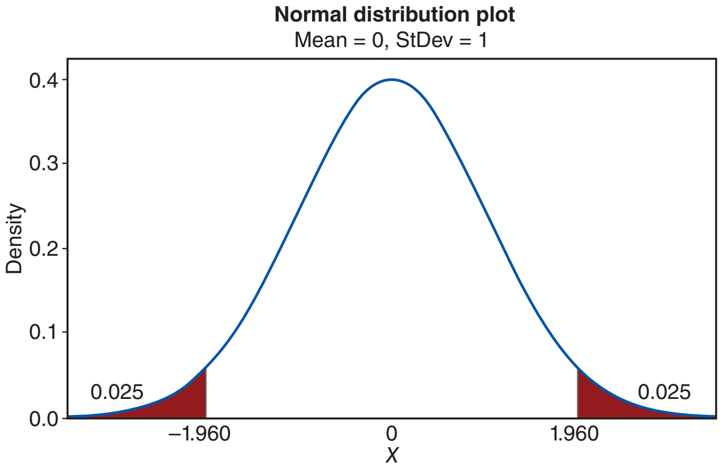


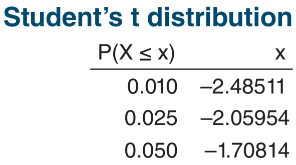
Figure 7.3.3 Probability density function of the $N(0, 1)$-distribution showing 0.025 areas under each tail.

MINITAB

(a) In order to determine the values of $t_{25,1-\alpha}$, we proceed as follows:

1. Enter the values 0.01, 0.025, and 0.05 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **t . . . .**
3. In the dialog box that appears, click the circle next to **Inverse probability**.
4. Complete the boxes next to **Degrees of freedom** and **Input Column** and click **OK**.

The values of $t_{25,1-\alpha}$ will appear in the Session window, as shown below.



Student's t distribution with 25 DF

P(X ≤ x)	x
0.010	–2.48511
0.025	–2.05954
0.050	–1.70814

(b) Recall the result that $t_{n,\alpha} = -t_{n,1-\alpha}$. So for this part the values of $t_{25,\alpha}$ will be the same as in part (a) but with positive signs.

USING R

R has a built in t-distribution function 'qt(p, df, ncp, lower.tail = TRUE)', where p is the probability, df is the degrees of freedom, ncp is the noncentrality parameter, and lower.tail = TRUE gives the quantile corresponding to the left tail probability p. So, referring to Example 7.3.4, in part (a), we use lower.tail = TRUE, and in part (b), lower.tail = FALSE as we need lower and upper tail quantiles, respectively. Run following R code in R Console window to obtain required probabilities as shown below.

```
prob = c(0.01, 0.025, 0.05)
qt(prob, df=25, ncp=0, lower.tail = TRUE)
#R output
[1] -2.485107 -2.059539 -1.708141

qt(prob, df=25, ncp=0, lower.tail = FALSE)
#R output
[1] 2.485107 2.059539 1.708141
```

As an immediate application of Theorem 7.3.6, we have the following theorem:

Theorem 7.3.7 *Let $X_1, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Let the random variables $\bar{X}$ and S^2 be the sample average and sample variance. Then, the random variable*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{7.3.14}$$

has the student t-distribution with $(n-1)$ degrees of freedom.

To establish Theorem 7.3.7, we recall that $\bar{X}$ is distributed as normal with mean μ and variance σ^2/n, so that $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. From Theorems 7.3.4 and 7.3.5 we have that $(n-1)S^2/\sigma^2$ is distributed as chi-square with $(n-1)$ degrees of freedom, independent of the sample mean. From Theorem 7.3.6, we know that the random variable T in (7.3.14) is distributed as the student t with $(n-1)$ degrees of freedom. Summarizing, we may write

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} \sim \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}$$

Now suppose that $\bar{X}_1$ and $\bar{X}_2$ are sample averages of independent samples of size n_1 and n_2 from the normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Then, $\bar{X}_1$ and $\bar{X}_2$ are $N(\mu_1, \sigma_1^2/n_1)$ and $N(\mu_2, \sigma_2^2/n_2)$ random variables, respectively, and are independent. If we consider the linear combination $\bar{X}_1 - \bar{X}_2$, from Theorem 6.4.2, we have the result given below in Theorem 7.3.8.

Theorem 7.3.8 *Let $\bar{X}_1$ and $\bar{X}_2$ be sample averages of independent samples of size n_1 and n_2 from the normal population $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Then $\bar{X}_1 - \bar{X}_2$, has the normal distribution $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.*

We can now state the following important theorem, Theorem 7.3.9, given below.

Theorem 7.3.9 *Let $\bar{X}_1$ and $\bar{X}_2$ be sample averages and S_1^2 and S_2^2 be sample variances of independent samples of size n_1 and n_2 from the normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Then, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the random variable*

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \tag{7.3.15}$$

where $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ (pooled variance), has the Student t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

To prove Theorem 7.3.9, we note that if in Theorem 7.3.8 we put $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$U = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

has the normal distribution $N(0,1)$. Furthermore,

$$V = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2}$$

is the sum of two independent random variables having the chi-square distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively. Thus, V has a chi-square distribution with $(n_1-1)+(n_2-1) = n_1+n_2-2$ degrees of freedom. Therefore, by Theorem 7.3.6,

$$T = \frac{U}{\sqrt{V/(n_1+n_2-2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

has the Student t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

7.3.3 Snedecor's *F*-Distribution

Another important distribution in the theory of sampling from the normal distribution is the F-distribution, which is summarized in Theorem 7.3.10. But we start with the following definition:

Definition 7.3.4 Let X_1 and X_2 be two independent random variables having chi-square distribution with ν_1 and ν_2 degrees of freedom, respectively. Then, the random variable $F = \dfrac{X_1/\nu_1}{X_2/\nu_2}$ is said to be distributed as Snedecor's F-distribution with ν_1 and ν_2 degrees of freedom.

Theorem 7.3.10 *The probability density function of the Snedecor F-distribution with ν_1 and ν_2 degrees of freedom is*

$$h(f) = \begin{cases} \frac{\Gamma[(\nu_1+\nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} f^{(\nu_1/2)-1} \left(1 + \frac{\nu_1 f}{\nu_2}\right)^{-(\nu_1+\nu_2)/2}, & f > 0 \\ 0, & \textit{otherwise} \end{cases} \tag{7.3.16}$$

The derivation of the probability distribution in (7.3.16) is not included here but is available on the book website: www.wiley.com/college/gupta/statistics2e. The random variable F of the Definition 7.3.4 is sometimes referred to as the variance ratio, or Snedecor's F-variable, and is often denoted by F_{ν_1,ν_2}, where ν_1 and ν_2 are known as numerator and denominator degrees of freedom, respectively. At this point, we may add that $t_\nu^2 \sim F_{1,\nu}$, as the reader may easily verify.

The *mean* and *variance* of the random variable F whose distribution is given by (7.3.16) are

$$\mu = \frac{\nu_2}{\nu_2 - 2} \quad \text{provided that} \quad \nu_2 > 2 \tag{7.3.17}$$

$$\sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad \text{provided that} \quad \nu_2 > 4 \tag{7.3.18}$$

respectively. Figure 7.3.4 shows the shape of the p.d.f. of the F-distribution for various values of ν_1 and ν_2 degrees of freedom. The F random variable is nonnegative, and its distribution is right-skewed. The shape of the distribution changes as the degrees of freedom change.

As an immediate application, we have the following theorem:

Theorem 7.3.11 *Let $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ be two independent random samples from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Let S_1^2 and S_2^2 be the sample variances. Then, the random variable F defined as*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{7.3.19}$$

is distributed as F_{ν_1,ν_2} with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

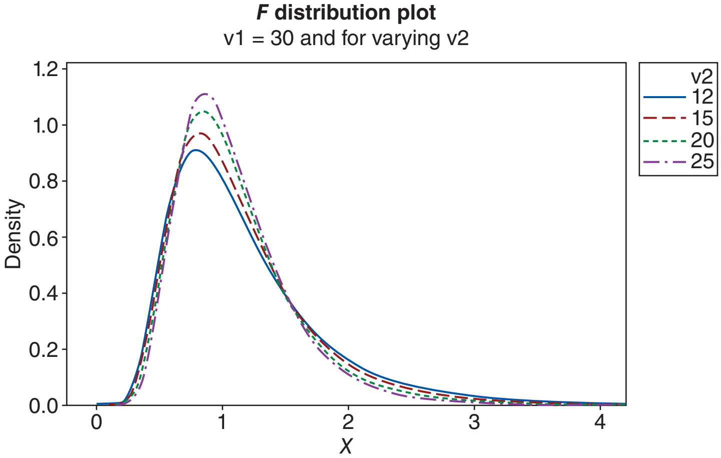


Figure 7.3.4 Probability density functions of F_{ν_1,ν_2} for various combinations of the degrees of freedom ν_1 (= 30) and ν_2 (= 12, 15, 20, 25).

The proof of this theorem follows directly by using the definition of random variable F_{ν_1,ν_2}. Recall from our earlier discussion, that random variables $(n_1 - 1)S_1^2/\sigma_1^2$ and $(n_1 - 1)S_2^2/\sigma_2^2$ are independently distributed as chi-square variables $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are the degrees of freedom. Values of $F_{\nu_1,\nu_2,\alpha}$ for which

$$P(F_{\nu_1,\nu_2} > F_{\nu_1,\nu_2,\alpha}) = \alpha$$

are given in Table A.7 for various values of ν_1, ν_2, and α. Note that to find values of $F_{\nu_1,\nu_2,1-\alpha}$ for which $P(F_{\nu_1,\nu_2} < F_{\nu_1,\nu_2,1-\alpha}) = \alpha$, we may proceed by using the easily proved relation

$$F_{\nu_1,\nu_2,1-\alpha} = \frac{1}{F_{\nu_2,\nu_1,\alpha}} \tag{7.3.20}$$

For example, the value of $F_{20,15,0.95} = F_{20,15,1-0.05}$ is given by

$$F_{20,15,0.95} = \frac{1}{F_{15,20,0.05}} = 0.4539$$

Example 7.3.5 (Using MINITAB and R to find probabilities of F-distribution) *Using MINITAB and R determine the following:*

(a) $F_{16,24,1-\alpha}$ *for* $\alpha = 0.01, 0.025, 0.05,$ *where* $F_{16,24,1-\alpha}$ *is such that* $P(F_{16,24} \leq F_{16,24,1-\alpha}) = \alpha.$

(b) $F_{16,24,\beta}$ *for* $\beta = 0.01, 0.025, 0.05,$ *where* $F_{16,24,\beta}$ *is such that* $P(F_{16,24} \geq F_{16,24,\beta}) = \beta.$

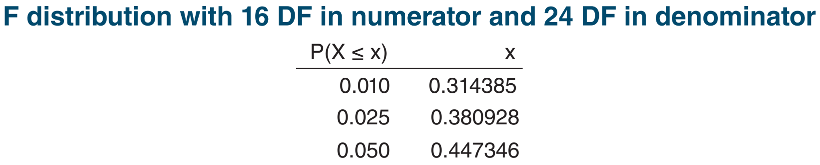
MINITAB

(a) In order to determine the values of $F_{16,24,1-\alpha}$, we proceed as follows:

1. Enter the values 0.01, 0.025, and 0.05 in column C1.
2. From the Menu bar, select **Calc** > **Probability Distribution** > **F . . . .**

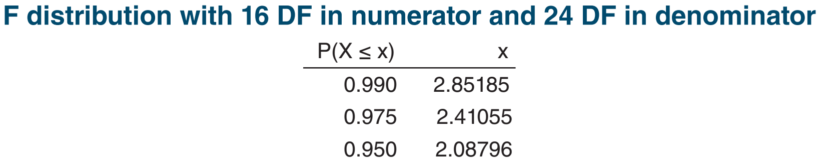
3. In the dialog box that appears, click the circle next to **Inverse cumulative probability**.
4. Complete the boxes next to **Numerator Degrees of freedom**, **Denominator Degrees of freedom**, and **Input Column**, and click **OK**.

The values of $F_{16,24,1-\alpha}$ will appear in the Session window as shown below:



```
F distribution with 16 DF in numerator and 24 DF in denominator

P(X ≤ x)          x
    0.010  0.314385
    0.025  0.380928
    0.050  0.447346
```

Note that MINITAB determines points that give lower tail areas under the distribution. Certain desired values under the right or upper tail are determined as follows: (b) Here, to determine the values of $F_{16,24,\beta}$ such that $P(F_{16,24} \geq F_{16,24,\beta}) = \beta$ or $P(F_{16,24} \leq F_{16,24,\beta}) = 1 - \beta$, accordingly we enter the values in column C1 as $(1 - 0.01, 1 - 0.025, 1 - 0.05) = (0.99, 0.975, 0.95)$ and proceed in the same manner as in part (a). We obtain the probabilities in the Session window as shown below:



```
F distribution with 16 DF in numerator and 24 DF in denominator

P(X ≤ x)        x
   0.990  2.85185
   0.975  2.41055
   0.950  2.08796
```

Note that the values of $F_{16,24,\alpha}$ are not given in F-tables.

USING R

R has a built in F-distribution function 'qf(p, df1, df2, ncp, lower.tail = TRUE)', where p is the probability, df1 and df2 are numerator and denominator degrees of freedoms, ncp is the noncentrality parameter, and lower.tail = TRUE gives the quantile corresponding to the left tail probability p. So, referring to Example 7.3.5, in part (a) we use lower.tail = TRUE and in part (b) lower.tail = FALSE as we need lower and upper tail quantiles, respectively. Run following R code in the R Console window to obtain required probabilities as shown below.

```
prob = c(0.01, 0.025, 0.05)
#For part (a)
qf(prob, df1=16, df2=24, ncp=0, lower.tail = TRUE)
#R output
[1] 0.3143853 0.3809283 0.4473461

#For part (b)
qf(prob, df1=16, df2=24, ncp=0, lower.tail = FALSE)
#R output
[1] 2.851852 2.410548 2.087963
```

PRACTICE PROBLEMS FOR SECTION 7.3

1. If X is a chi-square random variable with 15 degrees of freedom, find the value of x such that (a) $P(X \geq x) = 0.05$, (b) $P(X \geq x) = 0.975$, (c) $P(X \leq x) = 0.025$, (d) $P(X \geq x) = 0.95$, (e) $P(X \leq x) = 0.05$
2. Use Table A.6 to find the following values of upper percent points of various t_m-distributions: (a) $t_{18,0.025}$, (b) $t_{20,0.05}$, (c) $t_{15,0.01}$, (d) $t_{10,0.10}$, (d) $t_{12,0.005}$.
3. Use Table A.7 to find the following values of upper percent points of various F-distributions: (a) $F_{6,8,0.05}$, (b) $F_{8,10,0.01}$, (c) $F_{6,10,0.05}$, (d) $F_{10,11,0.025}$
4. Use Table A.7 to find the following values: (a) $F_{10,12,0.95}$, (b) $F_{8,10,0.975}$, (c) $F_{15,20,0.95}$, (d) $F_{20,15,0.99}$. Hint: Use the formula $F_{m,n,1-\alpha} = 1/F_{n,m,\alpha}$.
5. Use MINITAB, R, or JMP to do the Problems 1, 2, 3, and 4 above.
6. Suppose that the random variable T has the Student t-distribution with 24 degrees of freedom. Find the value of t such that (a) $P(-1.318 < T < t) = 0.80$, (b) $P(-1.711 < T < t) = 0.85$, (c) $P(-2.064 < T < t) = 0.875$.
7. Find the value of x such that (a) $P(3.247 < \chi^2_{10} < x) = 0.95$, (b) $P(8.260 < \chi^2_{20} < x) = 0.965$, (c) $P(13.120 < \chi^2_{25} < x) = 0.95$.

7.4 ORDER STATISTICS

In this section, we shall consider probability distributions of statistics that are obtained if one orders the n elements of a sample of n independent observations from least to greatest, and if sampling is done on a *continuous* random variable X whose p.d.f. is $f(x)$. Suppose we let $X_1, \ldots, X_n$ be a random sample of n independent observations from a population having continuous p.d.f. $f(x)$. We note that since X is a continuous random variable, the probability of X assuming a specific value is 0. In fact, by a straightforward conditional probability argument, we can show that for any two of $(X_1, \ldots, X_n)$, the probability of their having the same value is zero.

Consider then the observations $(X_1, \ldots, X_n)$ from a population having a p.d.f. $f(x)$. Let

$$\begin{aligned}
X_{(1)} &= \text{smallest of } (X_1, \ldots, X_n),\\
X_{(2)} &= \text{second smallest of } (X_1, \ldots, X_n),\\
&\vdots\\
X_{(k)} &= \text{kth smallest of } (X_1, \ldots, X_n),\\
&\vdots\\
X_{(n)} &= \text{largest } (X_1, \ldots, X_n).
\end{aligned}$$

Note that $X_{(1)} < X_{(2)} < \cdots < X_{(k)} < \cdots < X_{(n)}$. The quantities $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are random variables and are called the *order statistics* of the sample. $X_{(1)}$ is called the *smallest* element in the sample, $X_{(k)}$ the kth-order statistic; $X_{(m+1)}$ is the sample median when the sample size is odd, say $n = (2m + 1)$, and $X_{(n)}$ the largest; $R = X_{(n)} - X_{(1)}$ is called the sample range.

7.4.1 Distribution of the Largest Element in a Sample

As we have just stated, $X_{(n)}$ is the largest element in the sample $X_1, \ldots, X_n$. If the sample is drawn from a population having p.d.f. $f(x)$, let $F(x)$ be the cumulative density function (c.d.f.) of the population defined by

$$F(x) = \int_{-\infty}^{x} f(u)du = P(X \leq x) \tag{7.4.1}$$

Then, the c.d.f. of $X_{(n)}$ is given by

$$\begin{aligned} P(X_{(n)} \leq x) &= P(X_1 \leq x, \ldots, X_n \leq x) \\ &= [F(x)]^n \end{aligned} \tag{7.4.2}$$

because the X_i's are independent and $P(X_i \leq x) = F(x)$ for $i = 1, 2, \ldots, n$. If we denote the c.d.f. of the largest value by $G(x)$, we have

$$\boxed{G(x) = [F(x)]^n \qquad (7.4.3)}$$

The above result says that if we take a random sample of n elements from a population whose p.d.f. is $f(x)$ [or whose c.d.f. is $F(x)$], then the c.d.f. $G(x)$ of the largest element in the sample, denoted by X, is given by (7.4.3).

If we denote the p.d.f. of the largest element by $g_{X_{(n)}}(x)$, we have

$$g_{X_{(n)}}(x) = \frac{d}{dx} G(x) = n[F(x)]^{n-1} f(x). \tag{7.4.4}$$

Example 7.4.1 (Distribution of Last Bulb to Fail) *Suppose the mortality of a certain type of mass-produced light bulbs is such that a bulb of this type, taken at random from production, burns out in time T. Further, suppose that T is distributed as exponential with parameter λ, so that the p.d.f. of T is given by*

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t > 0 \\ 0, & t \leq 0 \end{cases} \tag{7.4.5}$$

where λ is some positive constant. If n bulbs of this type are taken at random, let their lives be $T_1, \ldots, T_n$. If the order statistics are $T_{(1)}, \ldots, T_{(n)}$, then $T_{(n)}$ is the life of the last bulb to burn out. We wish to determine the p.d.f. of $T_{(n)}$.

Solution: To solve this problem, we may think of a population of bulbs whose p.d.f. of length of life is given by equation (7.4.5), we first determine that the c.d.f. of T is given by

$$F(t) = \int_{-\infty}^{t} f(t)dt = \int_{0}^{t} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t} \tag{7.4.6}$$

Applying equation (7.4.4), we therefore have as the p.d.f. of $T_{(n)}$,

$$g_{T_n}(t) = \begin{cases} n\lambda(1-e^{-\lambda t})^{n-1}e^{-\lambda t}, & t > 0 \\ 0, & t \leq 0 \end{cases} \tag{7.4.7}$$

In other words, the probability that the last bulb to burn out expires during the time interval $(t, t+dt)$ is given by $g(t)dt$, where

$$g(t)dt = n\lambda(1-e^{-\lambda t})^{n-1}e^{-\lambda t}dt \tag{7.4.8}$$

7.4.2 Distribution of the Smallest Element in a Sample

We now wish to find the expression for the c.d.f, of the smallest element $X_{(1)}$ in the sample $X_1, \ldots, X_n$. That is, we want to determine $P(X_{(1)} \leq x)$ as a function of x.

Denoting this function by $G(x)$, we have

$$\begin{aligned} G(x) &= P(X_{(1)} \leq x) \\ &= 1 - P(X_{(1)} > x) \end{aligned} \tag{7.4.9}$$

But

$$\begin{aligned} P(X_{(1)} > x) &= P(X_1, \ldots, X_n \text{ are all} > x) \\ &= [1 - F(x)]^n \end{aligned} \tag{7.4.10}$$

because the X_i's are independent and $P(X_i > x) = 1 - F(x); i = 1, 2, \ldots, n$. Therefore, the c.d.f. $G(x)$ of the smallest element in the sample is given by

$$G(x) = 1 - [1 - F(x)]^n \tag{7.4.11}$$

The p.d.f., say $g(x)$, of the smallest element in the sample is therefore obtained by taking the derivative of the right-hand side of (7.4.11) with respect to x. We thus find

$$g(x) = n[1 - F(x)]^{n-1}f(x)$$

That is, the p.d.f. of $X_{(1)}$, is given by

$$g_{X_{(1)}}(x) = n[1 - F(x)]^{n-1}f(x) \tag{7.4.12}$$

Example 7.4.2 (Probability Distribution of the Weakest Link of a Chain) *Suppose links of a certain type used for making chains are such that the population of individual links has breaking strengths X with p.d.f.*

$$f(x) = \begin{cases} \frac{(m+1)(m+2)}{c^{m+2}}x^m(c-x), & 0 < x < c \\ 0, & \text{otherwise} \end{cases} \tag{7.4.13}$$

where c *and* m *are certain positive constants. If a chain is made up of n links of this type taken at random from the population of links, what is the probability distribution of the breaking strength of the chain?*

Solution: Since the breaking strength of a chain is equal to the breaking strength of its weakest link, the problem reduces to finding the p.d.f. of the smallest element $X_{(1)}$ in a sample of size n from the p.d.f. $f(x)$ given in (7.4.13).

First, we find the c.d.f. $F(x)$ of breaking strengths of individual links by performing the following integration:

$$F(x) = \int_{-\infty}^{x} f(u)du = \frac{(m+1)(m+2)}{c^{m+2}} \int_{0}^{x} u^m (c-u)du \tag{7.4.14}$$

that is,

$$F(x) = (m+2)\Big(\frac{x}{c}\Big)^{m+1} - (m+1)\Big(\frac{x}{c}\Big)^{m+2} \tag{7.4.15}$$

With the use of equations (7.4.12) and (7.4.13), we obtain the p.d.f. of the breaking strength X of an n-link chain made from a random sample of n of these links;

$$g(x) = n\frac{(m+1)(m+2)x^m}{c^{m+2}} \times \left[1-(m+2)\Big(\frac{x}{c}\Big)^{m+1} + (m+1)\Big(\frac{x}{c}\Big)^{m+2}\right]^{n-1}(c-x) \tag{7.4.16}$$

for $0 < x < c$, and $g(x) = 0$, otherwise.

7.4.3 Distribution of the Median of a Sample and of the *k*

Suppose we have a sample of $2m + 1$ elements $X_1, \ldots, X_{2m+1}$ from a population having p.d.f. $f(x)$ [and c.d.f. $F(x)$]. If we form the order statistics $X_{(1)}, \ldots, X_{(2m+1)}$ of the sample, then $X_{(m+1)}$ is called the sample median. We want to determine the probability distribution function for the median. Let us divide the *x-axis* into the following three disjoint intervals:

$$\begin{aligned} I_1 &= (-\infty, x] \\ I_2 &= (x, x+dx] \\ I_3 &= (x+dx, +\infty) \end{aligned} \tag{7.4.17}$$

Then, the probabilities $p_1, p_2, and\ p_3$ that an element X drawn from the population with p.d.f. $f(x)$ will lie in the intervals $I_1, I_2, and\ I_3$ are given, respectively, by

$$\begin{aligned} p_1 &= F(x) \\ p_2 &= F(x+dx) - F(x) \\ p_3 &= 1 - F(x+dx) \end{aligned} \tag{7.4.18}$$

respectively.

If we take a sample of size $2m + 1$ from the population with p.d.f. $f(x)$, the median of the sample will lie in $(x, x + dx)$ if, and only if, m sample elements fall in $I_1 = (-\infty, x]$, one sample element falls in $I_2 = (x, x+dx]$, and m sample elements fall in $I_3 = (x+dx, +\infty)$. The probability that all of this occurs is obtained by applying the multinomial probability distribution discussion in Section 4.7. This gives

$$\frac{(2m+1)!}{(m!)^2}(p_1)^m(p_2)^1(p_3)^m \tag{7.4.19}$$

But substituting the values of $p_1, p_2, and\ p_3$ from (7.4.18) into (7.4.19), we obtain

$$\frac{(2m+1)!}{(m!)^2}F^m(x)[F(x+dx)-F(x)][1-F(x+dx)]^m \tag{7.4.20}$$

Now, we may write

$$F(x+dx) = F(x) + f(x)dx \tag{7.4.21}$$

Substituting this expression into (7.4.20), we find that (ignoring terms of order $(dx)^2$ and higher)

$$P(x < X_{(m+1)} < x+dx) = \frac{(2m+1)!}{(m!)^2}F^m(x)[1-F(x)]^m f(x)dx \tag{7.4.22}$$

The p.d.f. $g(x)$ of the median is the coefficient of dx on the right-hand side of (7.4.22), and the probability that the sample median $X_{(m+1)}$ falls in interval $(x, x+dx)$ is given by

$$g_{X_{(m+1)}}(x)dx = \frac{(2m+1)!}{(m!)^2}F^m(x)[1-F(x)]^m f(x)dx. \tag{7.4.23}$$

We note that the sample space of the median $X_{(m+1)}$ is the same as the sample space of X, where X has the (population) c.d.f. $F(x)$.

Example 7.4.3 (Probability Distribution of Median) *Suppose* $2m+1$ *points are taken "'at random" on the interval (0,1). What is the probability that the median of the* $2m+1$ *points falls in* $(x, x+dx)$*?*

In this example, the p.d.f. of a point X taken at random on (0,1) is defined as

$$f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{for all other values of } x \end{cases}$$

Then,

$$F(x) = \begin{cases} 0, & x \le 0, \\ x, & 0 < x < 1, \\ 1, & x \ge 1 \end{cases}$$

Therefore, the p.d.f. $g_{X_{(m+1)}}(x)$ of the median in a sample of $2m+1$ points is given by

$$g_{X_{(m+1)}}(x) = \frac{(2m+1)!}{(m!)^2}x^m[1-x]^m, \quad if\ \ 0 < x < 1,$$

and zero otherwise. Hence, the probability that the median of the $2m+1$ points falls in $(x, x+dx)$ is given by

$$g_{X_{(m+1)}}(x)dx = \frac{(2m+1)!}{(m!)^2} x^m [1-x]^m dx$$

More generally, if we have a sample of n elements, say $X_1, \ldots, X_n$, from a population having p.d.f. $f(x)$ and if $X_{(k)}$ is the kth-order statistic of the sample (the kth smallest of $X_1, \ldots, X_n$), then we can show, as in the case of the median, that

$$P(x < X_{(k)} < x+dx) = \frac{n!}{(k-1)!(n-k)!} F^{k-1}(x)[1-F(x)]^{n-k} f(x)dx \tag{7.4.24}$$

Therefore, the p.d.f. of the kth-order statistic of the sample is given by

$$g_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F^{k-1}(x)[1-F(x)]^{n-k} f(x) \tag{7.4.25}$$

Note that the functional form of the p.d.f. on the right-hand side of (7.4.25) reduces to that on the right-hand side of (7.4.12) if k=1, and to that on the right of (7.4.4) if $k = n$, as one would expect, since in these two cases, the kth-order statistic $X_{(k)}$ becomes the smallest element $X_{(1)}$ and the largest element $X_{(n)}$, respectively.

Example 7.4.4 (Distribution of the kth order statistic) *If n points $X_1, \ldots, X_n$ are taken "at random" on the interval (0, 1) what is the p.d.f. of the* k*th order statistic $X_{(k)}$?*

Using (7.4.25), the p.d.f. of $X_{(k)}$ is given by:

$$g_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}[1-x]^{n-k}, \quad \text{if } 0 < x < 1$$

since

$$F(x) = \int_0^x 1dx = x, \quad \text{if } 0 < x < 1$$

and zero otherwise. Thus,

$$F(x) = \begin{cases} 0, & \text{if } x \le 0, \\ x, & \text{if } 0 < x < 1 \\ 1, & \text{if } x \ge 1 \end{cases}$$

7.4.4 Other Uses of Order Statistics

The Range as an Estimate of σ in Normal Samples

Suppose a random variable X has the normal distribution with unknown standard deviation σ. If a sample of n independent observations is taken on X, then $R = X_{(n)} - X_{(1)}$ may be used as the basis for an estimate of σ. This estimate is not good for large n, but for small $n(n \le 10)$ is deemed to be adequate. The estimate $\hat{\sigma}$ is made using the formula

$$\hat{\sigma} = c(n)R \tag{7.4.26}$$

where $c(n)$ is tabulated in Table 7.4.1.

Table 7.4.1 Range coefficient for various sample sizes.

n	3	4	5	6	7	8	9	10
$c(n)$	0.591	0.486	0.43	0.395	0.37	0.351	0.337	0.325

PRACTICE PROBLEMS FOR SECTION 7.4

1. A continuous random variable, say X, has the uniform distribution function on $(0, 1)$ so that the p.d.f. of X is given by

$$f(x) = \begin{cases} 0, & x \leq 0 \\ 1, & 0 < x \leq 1 \\ 0, & x > 1 \end{cases}$$

If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics of n independent observations all having this distribution function, give the expression for the density $g(x)$ for

(a) The largest of these n observations.
(b) The smallest of these n observations.
(c) The rth smallest of these n observations.

2. If ten points are picked independently and at random on the interval $(0, 1)$:

(a) What is the probability that the point nearest 1 (i.e., the largest of the 10 numbers selected) will lie between 0.9 and 1.0?
(b) The probability is 1/2 that the point nearest 0 will exceed what number?

3. Assume that the cumulative distribution function of breaking strengths (in pounds) of links used in making a certain type of chain is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

where λ is a positive constant. What is the probability that a 100-link chain made from these links would have a breaking strength exceeding y pounds?

4. Suppose $F(x)$ is the fraction of objects in a very large lot having weights less than or equal to x pounds. If 10 objects are drawn at random from the lot:

(a) What is the probability that the heaviest of 10 objects chosen at random without replacement will have a weight less than or equal to u pounds?
(b) What is the probability that the lightest of the objects will have a weight less than or equal to v pounds?

5. The time, in minutes, taken by a manager of a company to drive from one plant to another is uniformly distributed over an interval [15, 30]. Let $X_1, X_2, \ldots, X_n$ denote her driving times on n randomly selected days, and let $X_{(n)} = \text{Max}(X_1, X_2, \cdots, X_n)$. Determine

(a) The probability density function of $X_{(n)}$.
(b) The mean of $X_{(n)}$.

Note: Here, $f(x) = 1/15$ if $15 \leq x \leq 30$, and 0 otherwise.

6. The lifetime, in years, $X_1, X_2, \cdots, X_n$ of n randomly selected power steering pumps manufactured by a subsidiary of a car company is exponentially distributed with mean $1/\lambda$. Find the probability density function of $X_{(1)} = \text{Min}(X_1, X_2, \cdots, X_n)$, and find its mean and variance.

7. In Problem 5, assume that $n = 21$.
   (a) Find the probability density function of the median time taken by the manager to drive from one plant to another.
   (b) Find the expected value of $X_{(21)}$.
   (c) Find the expected value of $X_{(11)}$, the median.
8. Consider a system of n identical components operating independently. Suppose the lifetime, in months, is exponentially distributed with mean $1/\lambda$. These components are installed in series, so that the system fails as soon as the first component fails. Find the probability density function of the life of the system, and then, find its mean and variance.

7.5 USING JMP

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. The times taken by all students of a large university to complete a calculus test are distributed having a mean of 120 minutes and a standard deviation of 10 minutes. Calculate the approximate probability that the average time taken to complete their test by a random sample of 36 students will be (a) more than 122 minutes, (b) less than 115 minutes, (c) between 116 and 123 minutes.

2. Suppose that in a certain country, the ages of women at the time of death are distributed with mean 70 years and standard deviation 4 years. Find the approximate probability that the average age of a randomly selected group of 36 women will be (a) more than 75 years, (b) less than 70 years, (c) between 70 and 80 years.

3. A manufacturer of car batteries finds that 80% of its batteries last more than five years without any maintenance. Suppose that the manufacturer took a random sample of 500 persons from those who bought those batteries and recorded the lifetimes of their batteries.
   (a) Find the sampling distribution of $\hat{p}$, the proportion of batteries that lasted more than five years without any maintenance.
   (b) Find the probability that $\hat{p}$, the sample proportion, is at least 75%.

4. It is believed that the median annual starting salary of a fresh engineering graduate is $40,000. If we take a random sample of 100 recent engineering graduates and record their starting salary, then
   (a) Find the sampling distribution of $\hat{p}$, the proportion of fresh engineering graduates who started with an annual salary of less than $40,000.
   (b) Find the approximate probability that at least 60% of the engineering graduates started with an annual salary of more than $40,000.

5. Suppose that $F(x)$ is the fraction of bricks in a very large lot having crushing strengths of x psi or less. If 100 such bricks are drawn at random from the lot:
   (a) What is the probability that the crushing strengths of all 100 bricks exceed x psi?

(b) What is the probability that the weakest brick in the sample has a crushing strength in the interval $(x, x + dx)$?

6. Suppose that $n = 2m + 1$ observations are taken at random from a population with probability density function

$$f(x) = \begin{cases} 0, & x \leq a \\ \frac{1}{b-a}, & a < x \leq b \\ 0, & x > b \end{cases}$$

Find the distribution of the median of the observations and find its mean and variance. What is the probability that the median will exceed $a + (b - a)/4$?

7. Suppose that the diameter of ball bearings used in heavy equipment are manufactured in a certain plant and are normally distributed with mean 1.20 cm and a standard deviation 0.05 cm. What is the probability that the average diameter of a sample of size 25 will be

(a) Between 1.18 and 1.22 cm?
(b) Between 1.19 and 1.215 cm?

8. A sample of n observations is taken at random from a population with p.d.f.

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Find the p.d.f. of the smallest observation. What are its mean and variance? What is its c.d.f.?

9. If $X_{(1)}, \ldots, X_{(n)}$ are the order statistics of a sample of size n from a population having a continuous c.d.f. $F(x)$ and p.d.f. $f(x)$, show that $F(x_{(n)})$ has mean $n/(n + 1)$ and variance $n/[(n + 1)^2(n + 2)]$.

10. In Problem 9, show that for $1 \leq k \leq n$, the mean and variance of $F(x_{(k)})$ are respectively

$$\frac{k}{n+1} \quad \text{and} \quad \frac{k(n-k+1)}{(n+1)^2(n+2)}$$

11. Suppose that $X_{(1)}, \ldots, X_{(n)}$ are the order statistics of a sample from a population having the rectangular distribution with p.d.f.

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{\theta}, & 0 < x \leq \theta \\ 0, & x > \theta \end{cases}$$

where θ is an unknown parameter. Show that for $0 < \gamma < 1$,

$$P\left(X_{(n)} \leq \theta \leq \frac{X_{(n)}}{\sqrt[n]{1-\gamma}}\right) = \gamma$$

(Note: $\sqrt[n]{1-\gamma} = (1-\gamma)^{1/n}$).

12. Find, using MINITAB/R/JMP, the value of x such that
    (a) $P(5 \leq \chi^2_{16} \leq x) = 0.95$
    (b) $P(10 \leq \chi^2_{20} \leq x) = 0.90$
    (c) $P(0.5 \leq F_{20,24} \leq x) = 0.90$
    (d) $P(0.4 \leq F_{12,15} \leq x) = 0.85$

13. Let $X_1, \ldots, X_{16}$ and $Y_1, \ldots, Y_{13}$ be two independent random samples from two normal populations with equal variances. Show that the p.d.f. of S^2_x/S^2_y is distributed as Snedecor's $F_{15,12}$.

14. Refer to Problem 13. Using MINITAB, R, or JMP, find the probabilities: (a) $P(S^2_x/S^2_y > 3.5)$, (b) $P(2.0 < S^2_x/S^2_y < 3.5)$.

15. Suppose that the total cholesterol levels of the US male population between 50 and 70 years of age are normally distributed with mean 170 mg/dL and standard deviation 12 mg/dL. Let $X_1, \ldots, X_{21}$ be the cholesterol levels of a random sample of 21 US males between the ages of 50 and 70 years. What can be said about the p.d.f. of S^2. Find the mean and variance of S^2.

16. In Problem 15, using MINITAB, R, or JMP, find the probabilities: (a) $P(S^2 > 100)$, (b) $P(S^2 > 130)$, (c) $P(S^2 > 140)$.

17. A mechanical system has three components in series, so the system will fail when at least one of the components fails. The random variables X_1, X_2, and X_3 represent the lifetime of these components. Suppose that the $X_i (i = 1, 2, 3)$ are independently and identically distributed as exponential with parameter $\lambda = 0.1$. Let the random variable T denote the lifetime of the system. Find the p.d.f. of T, and then find the probability $P(T > 10)$.

18. Suppose that in Problem 17, the components are in parallel, so that the system will fail only when all the components fail. Find the p.d.f. of T, and then find the probability $P(T > 15)$.

19. Repeat Problems 17 and 18, by supposing that the lifetimes of the three components are independently and identically distributed as Weibull with $\alpha = 2, \beta = 0.5$.

20. Seven engineers in a manufacturing company are working on a project. Let random variables $T_1, \ldots, T_7$ denote the time (in hours) needed by the engineers to finish the project. Suppose that $T_1, \ldots, T_7$ are independently and identically distributed by the uniform distribution over an interval $[0, 1]$.
    (a) Find the distribution of the sample median $T_{(4)}$.
    (b) Find the probabilities (i) $P(T_{(4)} > 0.9)$, (ii) $P(0.6 < T_{(4)} < 0.9)$, where $T_{(4)}$ is the fourth-order statistic of $(T_{(1)}, \ldots, T_{(7)})$.

Chapter 8

ESTIMATION OF POPULATION PARAMETERS

The focus of this chapter is the development of methods for finding point and interval estimators of population parameters.

Topics Covered

- Point estimators for the population mean and variance
- Interval estimators for the mean of a normal population
- Interval estimators for the difference of means of two normal populations
- Interval estimators for the variance of a normal population
- Interval estimators for the ratio of variances of two normal populations
- Estimation methods: method of moments and the method of maximum likelihood
- Point and interval estimators for the binomial parameter
- Prediction: estimation of a future observation

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the role of estimation in applied statistics.
- Determine the point and interval estimates for parameters of various discrete and continuous distributions.
- Understand the concept of a confidence coefficient and interpret confidence intervals.
- Understand the distinction between statistical and practical significance.
- Apply statistical packages MINITAB, R, and JMP to determine confidence intervals of parameters of various distributions.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

8.1 INTRODUCTION

We often encounter statistical problems of the following type: we have a large lot or population of objects such that if a measurement was made on each object, we would have a distribution of these measurements. Since these measurements have not been made, or if it is not possible to make measurements on all the objects in the population, this distribution is, of course, unknown. About the best we can hope to do in practice is to estimate various *characteristics* (commonly referred to as *parameters*) of this distribution from the information contained in measurements made in a random sample of objects from the lot or population. For instance, if we wish to estimate the mean and variance of the population distribution, it turns out that we can use the sample average and sample variance as estimators for these quantities. If we want to estimate the median of the population, we can use the sample median. Of course, other parameters of the population distribution such as, standard deviation, or population proportion can be estimated. Problems of this type are the focus of this chapter.

We consider two kinds of estimators for population parameters, namely *point estimators* and *interval estimators.* More specifically, suppose that $(X_1, \ldots, X_n)$ is a random sample from a population whose distribution has a parameter of interest, say θ. If $\hat{\theta} = \varphi(X_1, \ldots, X_n)$ is a (single-valued) function of $X_1, \ldots, X_n$ so that $\hat{\theta}$ is itself a random variable, we refer to θ as a *statistic.* Furthermore, if

$$E(\varphi(X_1, \ldots, X_n)) = E(\hat{\theta}) = \theta \tag{8.1.1}$$

we say that $\hat{\theta}$ is an *unbiased estimator* of θ. The statistic $\hat{\theta}$ (read as θ-"hat") is usually referred to as a *point estimator* for the parameter θ. If $\hat{\theta}_l = \varphi_l(X_1, \ldots, X_n)$ and $\hat{\theta}_u = \varphi_u(X_1, \ldots, X_n)$ are two statistics such that

$$P(\hat{\theta}_l < \theta < \hat{\theta}_u) = 1 - \alpha \tag{8.1.2}$$

we say that the random interval $(\hat{\theta}_l, \hat{\theta}_u)$ is a $100(1-\alpha)\%$ *confidence interval* for the parameter θ. The pair of statistics $(\hat{\theta}_l, \hat{\theta}_u)$ is sometimes referred to as an *interval estimator* for θ. The endpoints ($\hat{\theta}_l$ and $\hat{\theta}_u$) of the confidence interval $(\hat{\theta}_l, \hat{\theta}_u)$ are sometimes called the $100(1-\alpha)\%$ *confidence limits* of θ, while ($\hat{\theta}_l$ and $\hat{\theta}_u$) are usually referred to as the lower confidence limit (LCL) and the upper confidence limit (UCL), respectively. The probability $1-\alpha$ in Equation (8.1.2) is called the *confidence coefficient*, or while $100(1-\alpha)\%$, is called the confidence level. Also, the difference $\hat{\theta}_u - \hat{\theta}_l$ is referred to as the *width* of the confidence interval.

8.2 POINT ESTIMATORS FOR THE POPULATION MEAN AND VARIANCE

As important examples of point estimators, we consider the most commonly used point estimators for the mean and variance of a population. Suppose that it is a population in which the variable (or measurement) X is a continuous random variable and has a probability density function (p.d.f.) $f(x)$. As we have seen in Chapter 5, the population *mean* μ and *variance* σ^2 are defined as follows:

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \tag{8.2.1}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \tag{8.2.2}$$

For the case of a population in which X is a discrete random variable and has a probability function (p.f.) $p(x)$, we similarly define μ and σ^2 by using the operation of summation rather than integration. If the population distribution is unknown, then so are μ and σ^2. The basic question is this: How can we estimate μ and σ^2 from a random sample $(X_1, \ldots, X_n)$ drawn from the population? There are many ways of devising estimators. A simple point estimator for μ, and the most widely used, is the sample average $\bar{X}$. The sample average $\bar{X}$ is a random variable that has its own distribution and therefore has its own mean and variance (see Chapter 7):

$$E(\bar{X}) = \mu \tag{8.2.3}$$

and

$$Var(\bar{X}) = \sigma^2/n \tag{8.2.4}$$

It should be noted that the statistic $\bar{X}$ has a unique value for any given sample and can be represented as a point on a μ axis. If we consider an indefinitely large number of samples from this population, each of size n, then Equation (8.2.3) essentially states that if we were to average the $\bar{X}$'s of these samples, their average would be equal to μ. Then, we say that $\bar{X}$ is an unbiased (point) estimator for μ. Furthermore, we note from Equation (8.2.4) that if we were to determine the variance of all these $\bar{X}$'s, the result would be σ^2/n, which gives some indication of how all these $\bar{X}$'s would be distributed around the value μ. Note particularly that the larger the value of n, the more closely these $\bar{X}$'s will cluster around μ. Thus, in a similar manner, the sample variance S^2, where

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

can be used as a point estimator for the population variance σ^2. The statistic S^2 is a random variable with its own distribution, and the mean of this distribution is σ^2. That is,

$$E(S^2) = \sigma^2 \tag{8.2.5}$$

and hence S^2, is an unbiased *point* estimator for σ^2. The derivations of the results in Equations (8.2.3), (8.2.4), and (8.2.5) have been given in Chapter 7. We summarize the above results in the following theorem:

Theorem 8.2.1 *If $\bar{X}$ and S^2 are determined from a random sample of size n from a population with unknown mean μ and unknown variance σ^2, then $\bar{X}$ is an unbiased (point) estimator for μ having variance $\sigma^2_{\bar{X}} = \sigma^2/n$. Furthermore, S^2 is an unbiased (point) estimator for σ^2.*

8.2.1 Properties of Point Estimators

There are various properties of a good point estimator that are often met, such as unbiasedness, minimum variance, efficient, consistent, and sufficient. The properties that we discuss here in some detail are:

- Unbiasedness
- Minimum variance

Let $f(x, \theta)$ be the p.d.f. of a population of interest with an unknown parameter θ, and let $X_1, \ldots, X_n$ be a random sample from the population of interest. Following our earlier discussion, let $\hat{\theta} = \varphi(X_1, \ldots, X_n)$ be a point estimator of the unknown parameter θ. Then, we have:

Definition 8.2.1 The point estimator $\hat{\theta} = \varphi(X_1, \ldots, X_n)$ is said to be an *unbiased estimator* of θ if and only if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, then $\hat{\theta}$ is *biased*, and the difference, $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$, is called the *bias* of $\hat{\theta}$.

From Theorem 8.2.1, we have that $\bar{X}$ and S^2 are unbiased estimators of the population mean μ and the population variance σ^2, respectively. Having said this, there still remains an important question to be answered: If the mean μ of a population is unknown, and if we take a random sample from this population and calculate the sample average $\bar{X}$, using it as an estimator of μ, how do we know how close our estimator $\bar{X}$ is to the true value of μ? The answer to this question depends on the population size and the sample size.

Let E be the *maximum absolute difference* between an estimator $\bar{X}$ and the true value of μ, and let $0 < \alpha < 1$. Then, if the population is either normal with no restriction on the sample size or a nonnormal infinite population and the sample size is large ($n \geq 30$), we say with probability $(1 - \alpha)$ that

$$\text{Margin of error} : E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{8.2.6}$$

assuming σ to be known. The quantity E is called the *margin of error* or *bound on the error of estimation.* Note that in practice, it is quite common to take $\alpha = 0.01, 0.05$, or 0.1 since this gives a reasonably high probability $(1 - \alpha)$ that the maximum error of estimation is equal to E.

The result in Equation (8.2.6) is still valid if the population is finite, and the sampling is done with replacement or if the sampling is done without replacement, but the sample size is less than 5% of the population size ($n < 0.05N$). If the population is finite and the sample size relative to the size of the population is greater than 5%, then we use an extra factor referred to as the *finite population correction factor* $\sqrt{(N - n)/(N - 1)}$, and the maximum absolute difference between the estimator $\bar{X}$ and the true value of μ is

$$\text{Margin of error : } E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{8.2.7}$$

where N and n are the population size and the sample size, respectively. If σ is *not known* in Equations (8.2.6) and (8.2.7), then in a large sample, it can be replaced with the sample standard deviation S.

Example 8.2.1 (Margin of error) *A manufacturing engineer wants to use the mean of a random sample of size* $n = 64$ *to estimate the mean length of the rods being manufactured. If it is known that* $\sigma = 0.5$ *cm, then find the margin of error with 95% probability.*

Solution: Since the sample size is large and assuming that the total number of rods manufactured at the given facility is quite large, it follows from Equation (8.2.6) that

$$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Here, $1 - \alpha = 0.95, \alpha/2 = 0.025, \sigma = 0.5, z_{0.025} = 1.96$, and $n = 64$, so we find

$$E = 1.96 \times \frac{0.5}{\sqrt{64}} = 0.1225 \text{ cm}$$

Note that the nonabsolute value of E is 0.1225 cm.

Another desirable property of an unbiased estimator is whether or not it is the minimum variance estimator. If it is, then it will result in an estimate that is closer to the true value of the parameter.

Definition 8.2.2 Consider a population having p.d.f. $f(x, \theta)$ with an unknown parameter θ. Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ be a class of unbiased estimators of θ. Then, an estimator $\hat{\theta}_i$ is said to be an unbiased minimum-variance (UMV) estimator of θ if the variance of $\hat{\theta}_i$ is less than or equal to the variance of any other unbiased estimator.

Definition 8.2.3 The mean-squared error (*MSE*) of an estimator $\hat{\theta}$ of θ is defined as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{8.2.8}$$

It is readily seen that we can write Equation (8.2.8) as

$$MSE(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2 \tag{8.2.8a}$$

Thus, if $\hat{\theta}$ is an unbiased estimator of θ, then the mean squared error is given by

$$MSE(\hat{\theta}) = Var(\hat{\theta}) \tag{8.2.9}$$

Sometimes a goal, when choosing an estimator, is to minimize the squared error. If the estimator is unbiased, then from Equation (8.2.8a), we see that the squared error is minimized, and the variance of the estimator is a minimum. An estimator with minimum variance is called the *most efficient estimator*.

Definition 8.2.4 Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of a population parameter θ. If the variance of $\hat{\theta}_1$ is smaller than the variance of $\hat{\theta}_2$, we say that the estimator $\hat{\theta}_1$ is a *more efficient estimator* of θ than $\hat{\theta}_2$.

Figure 8.2.1 shows that when $\hat{\theta}_1$ is a more efficient estimator of θ than $\hat{\theta}_2$, then the distribution of $\hat{\theta}_1$ is more clustered around θ than that of $\hat{\theta}_2$. Thus, the probability is greater that the estimator $\hat{\theta}_1$ is closer to θ than $\hat{\theta}_2$.

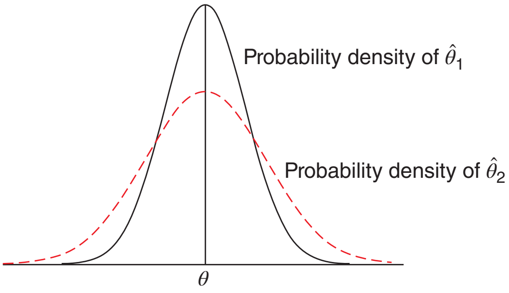


Figure 8.2.1 Probability distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$.

For example, let $X_1, \ldots, X_n$ be a random sample from an infinite population with symmetric distribution which has unknown mean μ and standard deviation σ, and let $\bar{X}$ and M_d be the sample average and the sample median, respectively. Then, both $\bar{X}$ and M_d are unbiased estimators of μ. However, for large samples from populations having symmetric continuous distributions, the standard error of the sample median is approximately $1.25 \times (\sigma/\sqrt{n})$, which is greater than the standard error of the sample average. Hence, for large samples from populations having symmetric continuous distributions, $\bar{X}$ is a better unbiased estimator of μ than the sample median.

Example 8.2.2 (Chemical batch data) *In order to evaluate a new catalyst in a chemical production process, a chemist used it in 30 batches. The final yield of the chemical in each batch is recorded as follows:*

72 74 71 78 84 80 79 75 77 76 74 78 88 78 70 72 84 82 80 75 73 76 78 84 83
85 81 79 76 72

(a) Find a point estimate of the mean yield of the chemical when the new catalyst is used.
(b) Find the standard error of the point estimator used in part (a).
(c) Find, with 95% probability, the margin of error.

Solution: (a) Since the sample size is large, all the results discussed previously are applicable to this problem. Thus, we have

$$\hat{\mu} = \bar{X} = (72 + 74 + 71 + 78 + \cdots + 72)/30 = 77.8$$

(b) To find the standard error of the point estimator used in (a), we first need to determine the sample standard deviation S, which is given by

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Now, substituting the values of n, X_i's, and $\bar{X}$ in the expression above, we have

$$S = 4.6416$$

so that the standard error of the point estimator $\bar{X}$ is estimated as

$$\frac{S}{\sqrt{n}} = \frac{4.6416}{\sqrt{30}} = 0.8474$$

(c) Now, we want to find the margin of error with 95% probability, $\alpha = 0.05$, and the population standard deviation σ is not known. Thus, substituting the value of $z_{\alpha/2} = z_{0.025} = 1.96, S = 4.6416$, and $n = 30$ into Equation (8.2.6), we find that the estimated margin of error is equal to $\pm 1.96(4.6416/\sqrt{30}) = \pm 1.6609$. We can summarize this by stating that the numerical value of E is 1.6609.

8.2.2 Methods of Finding Point Estimators

In the preceding section, we gave the point estimators of a population mean and population variance. In this section, we discuss two commonly used methods for finding point estimators: the *method of moments* and the *method of maximum likelihood.*

Method of Moments

The method of moments proposed by Karl Pearson is the oldest method for finding point estimators. It involves equating as many sample moments to the corresponding population moments as the number of unknown parameters and then solving the resulting equations for the unknown parameters.

Let $X_1, \ldots, X_n$ be a random sample. We define the rth population moment and rth sample moment as follows:

$$r\text{th population moment: } \mu'_r = E(X^r)$$

$$r\text{th sample moment: } m'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

Suppose now that $X_1, X_2, \ldots, X_n$ is a random sample from a population having probability distribution with unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then, their *moment estimators* $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are obtained by equating the first k population moments to the corresponding k sample moments. That is,

$$\begin{aligned} \mu'_1 &= m'_1 \\ \mu'_2 &= m'_2 \\ &\vdots \\ \mu'_k &= m'_k \end{aligned} \tag{8.2.10}$$

where $\mu'_r, r = 1, 2, \ldots, k$ are k functions of the unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$. Now solving Equations (8.2.10) for $\theta_1, \theta_2, \ldots, \theta_k$, we obtain the moment estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ for $\theta_1, \theta_2, \ldots, \theta_k$, respectively. The following examples illustrate this method.

Example 8.2.3 (Moment estimators) *Consider a random sample $X_1, \ldots, X_n$ from a population having the $N(\mu, \sigma^2)$ distribution. Find the moment estimators for μ and σ^2.*

Solution: We have two unknown parameters μ and σ^2. By equating two population moments with the corresponding two sample moments, we have

$$\mu'_1 = m'_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu'_2 = m'_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Since in this example, μ and σ^2 are the population mean and population variance, we have

$$E(X) = \mu'_1 = \mu \quad \text{and} \quad E(X^2) = \mu'_2 = \mu^2 + \sigma^2$$

These equations lead to the following two equations:

$$\mu = m'_1 \quad \text{and} \quad \mu^2 + \sigma^2 = m'_2$$

We solve these two equations for μ and σ^2 and obtain

Moment estimators for the population mean and the population variance are given by $\hat{\mu}$ and $\hat{\sigma}^2$, where

$$\hat{\mu} = m'_1 = \bar{X} \tag{8.2.11}$$

$$\hat{\sigma}^2 = m'_2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{8.2.11a}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{8.2.11b}$$

Note that for $\hat{\mu}$ and $\hat{\sigma}^2$ in Equations (8.2.11) and (8.2.11a), $\hat{\mu}$ is an unbiased estimator for μ, whereas $\hat{\sigma}^2$ is not an unbiased estimator for σ^2. This result leads us to conclude that the point estimators determined by using the method of moments may or may not be unbiased. Further, note that the moment estimator of σ is biased and given by Equation (8.2.11b).

Example 8.2.4 (Moment estimator for the mean of the Poisson) *Suppose that $X_1, \ldots, X_n$ is a random sample from a population with Poisson probability distribution with parameter λ. Find the moment estimator of λ.*

Solution: We have one unknown parameter λ. By equating one population moment with the corresponding one sample moment, we have

$$\mu_1' = m_1' = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Since in this example, λ is the population mean, we have

$$E(X) = \mu_1' = \lambda$$

This result leads us to the following:

$$\hat{\lambda} = m_1' = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Method of Maximum Likelihood

We now briefly describe a very widely used method of determining point estimators known as the method of *maximum likelihood estimation.* Suppose that $X_1, \ldots, X_n$ is a random sample on a random variable X whose probability function (in the discrete case) is $p(x, \theta)$, or X has a probability density function (in the continuous case), say $f(x, \theta)$, which depends on the population parameter θ. For example, the binomial distribution defined by Equation (4.6.2) depends on the parameter θ. The Poisson distribution Equation (4.8.1) depends on the parameter λ. A distribution function may depend, of course, on more than one parameter; for example, θ may stand for the pair of parameters (μ, σ), which is the case when X has the normal distribution $N(\mu, \sigma^2)$, having the probability density function defined in Equation (5.5.1).

Since $X_1, \ldots, X_n$ is a random sample from $f(x, \theta)$, the joint probability density function (or probability function) of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \times \cdots \times f(x_n; \theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{8.2.12}$$

which is usually denoted by $l(\theta|x_1, \ldots, x_n)$. The function $l(\theta|x_1, \ldots, x_n)$ is called the *likelihood function* of θ for the given sample $X_1, \ldots, X_n$.

Definition 8.2.5 The maximum likelihood estimator (MLE) $\hat{\theta}$ for θ is the value of θ (if it exists) which is such that

$$l(\hat{\theta}|x_1, \ldots, x_n) > l(\theta'|x_1, \ldots, x_n) \tag{8.2.13}$$

where θ' is any other possible value of θ. In other words, the MLE is the value of θ that maximizes the likelihood function $l(\theta|x_1, \ldots, x_n)$.

In applying Definition 8.2.5, it is usually more convenient to work with the natural logarithm of the likelihood function $l(\theta|x_1, \ldots, x_n)$ than with the likelihood function itself. We denote the natural logarithm of the likelihood function $l(\theta|x_1, \ldots, x_n)$ by $L(\theta)$. Further, it can be seen that $L(\theta)$ has its maximum at the same value of θ that $l(\theta|x_1, \ldots, x_n)$ does. Thus, the value of θ at which $L(\theta)$ is maximized is called the *maximum likelihood estimator* (MLE) $\hat{\theta}$ of θ. The MLE of θ is found by solving the following so-called normal equation:

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \tag{8.2.14}$$

We illustrate the method of maximum likelihood below with Examples 8.2.5 and 8.2.6.

Example 8.2.5 (Maximum likelihood estimate of the binomial parameter) *Find the maximum likelihood estimate of* p, *where* p *is the fraction of defectives of connecting rods for car engines produced by a mass-production process. We select a random sample of size* n *of connecting rods from the production line to observe the number of defectives.*

Solution: For a single connecting rod, let X be a random variable with value 1 if the rod is defective and 0 if the rod is nondefective. The probability function of X, say $p(x;p)$, is given by

$$p(x;p) = p^x(1-p)^{1-x}; \quad x = 0, 1 \tag{8.2.15}$$

In general, a random variable X that has p.f. given by Equation (8.2.15) is called a Bernoulli random variable. For a sample of n rods, the observations on X would be $(X_1, \ldots, X_n)$, and hence

$$l(p|x_1, \ldots, x_n) = p^T(1-p)^{n-T}$$

where $T = \sum_{i=1}^{n} X_i$, so that T is the number of defectives in the sample. Therefore,

$$L = T \ln p + (n-T)\ln(1-p)$$

Differentiating this with respect to p, we have

$$\frac{\partial L}{\partial p} = \frac{T}{p} - \frac{n-T}{1-p}$$

Setting this derivative equal to zero, we obtain a normal equation. Solving the normal equation for p and denoting the solution by $\hat{p}$, we have that the MLE for p is given by

Maximum likelihood estimator of the binomial parameter:

$$\hat{p} = \frac{T}{n} \tag{8.2.16}$$

In Section 8.7, we show that $\hat{p}$ is an unbiased estimator of p. Maximum likelihood estimators do not always have this latter property, as the next example shows.

Example 8.2.6 (Maximum likelihood estimators for the parameters of a normal distribution) *A random sample $X_1, \ldots, X_n$ of* n *independent observations is taken on* X*, where* X *is a random variable having the normal distribution $N(\mu, \sigma^2)$. Find the* MLEs *for μ and σ^2.*

Solution: The likelihood function is given by

$$\begin{aligned} l(\theta|x_1, \ldots, x_n) &= l(\mu, \sigma^2|x_1, \ldots, x_n) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned} \tag{8.2.17}$$

Hence,

$$L(\mu, \sigma^2|x_1, \ldots, x_n) = -n \ln\sqrt{2\pi} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{8.2.18}$$

To find the maximum of this function with respect to μ and σ^2, we differentiate partially with respect to μ and with respect to σ^2, obtaining

$$\begin{aligned} \frac{\partial L}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\ \frac{\partial L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned} \tag{8.2.19}$$

Setting these derivatives equal to zero, we obtain a set of equations called the *normal equations*. Solving the normal equations for μ and σ^2, we obtain the MLE $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2, respectively. These are

Maximum likelihood estimators of the mean and variance of a normal distribution:

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned} \tag{8.2.20}$$

We now show that $\hat{\mu} = \bar{X}$ is an unbiased estimator for μ, whereas $\hat{\sigma}^2$ is not an unbiased estimator of σ^2 . We have

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

that is, $\bar{X}$ is unbiased for μ. Further

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2\right) = \frac{\sigma^2}{n}E\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}\right) = \frac{\sigma^2}{n}E(\chi^2_{n-1}) = \frac{n-1}{n}\sigma^2$$

that is, $\hat{\sigma}^2$ is not unbiased for σ^2.

Using Equation (8.2.20), the reader can easily verify that the MLEs of the parameters μ and σ^2 in the lognormal distribution (see Section 5.9) are given by

Maximum likelihood estimators of the parameters of a lognormal distribution:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} \ln X_i \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2 \end{aligned} \qquad (8.2.21)$$

PRACTICE PROBLEMS FOR SECTION 8.2

1. Find a method of moments estimator of the mean of a Bernoulli distribution with parameter p.
2. The lengths of a random sample of 20 rods produced the following data:

12.2	9.5	13.2	13.9	9.5	9.5	11.9	9.2	11.0	10.4
9.9	12.8	10.5	11.9	12.3	10.0	8.7	6.2	10.0	11.2

   Determine the method of moments estimate of μ and σ^2, assuming that the rod lengths are normally distributed with mean μ and variance σ^2.
3. Let X_1, X_2, and X_3 be independent random variables with mean μ and variance σ^2. Suppose that $\hat{\mu}_1$ and $\hat{\mu}_2$ are two estimators of μ, where $\hat{\mu}_1 = 2X_1 - 2X_2 + X_3$ and $\hat{\mu}_2 = 2X_1 - 3X_2 + 2X_3$.
   (a) Show that both estimators are unbiased for μ.
   (b) Find the variance of each estimator and comment on which estimator is better.
4. Suppose that S_1^2 and S_2^2 are sample variances of two samples of n_1 and n_2 independent observations, respectively, from a population with mean μ and variance σ^2. Determine an unbiased estimator of σ^2 as a combination of S_1^2 and S_2^2.

5. The following data give the pull strength of 20 randomly selected solder joints on a circuit board. Assume that the pull strengths are normally distributed with mean μ and variance σ^2.

12	11	12	9	8	11	11	11	8	9
10	9	8	11	10	8	9	9	11	10

   (a) Determine the maximum likelihood estimate of the population mean of pull strengths of solder joints.
   (b) Determine the maximum likelihood estimate of population variance of pull strengths of solder joints.
6. Suppose that a random sample of size n is taken from a gamma distribution with parameters γ and λ. Find the method of moments estimators of γ and λ.
7. If $(X_1, \ldots, X_n)$ is a random sample of size n from a population having a Poisson distribution with unknown parameter λ, find the MLE for λ.
8. Events occur in time in such a way that the time interval between two successive events is a random variable t having the p.d.f. $\theta e^{-\theta t}$, $\theta > 0$. Suppose that observations are made of the n successive time intervals for $n + 1$ events, yielding $(t_1, \ldots, t_n)$. Assuming these time intervals to be independent, find the MLE for θ.
9. Suppose $(X_1, \ldots, X_n)$ is a random sample of n independent observations from a population having p.d.f.
$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$
Find the MLE of θ. If $\hat{\theta}$ is the MLE of θ, then show that $\hat{\theta}$ is not unbiased for θ.
10. Referring to Problem 9,
   (a) Show that $U = (n + 1)\hat{\theta}/n$ is unbiased for θ.
   (b) Find the variance of U.
   (c) Verify that $2\bar{X}$ is unbiased for θ. Find the variance of $2\bar{X}$.
   (d) Determine the ratio $Var(U)/Var(2\bar{X})$. Which of the unbiased estimators would you prefer?

8.3 INTERVAL ESTIMATORS FOR THE MEAN μ OF A NORMAL POPULATION

8.3.1 σ^2 Known

A general method to determine a confidence interval for an unknown parameter θ makes use of the so called *pivotal quantity.*

Definition 8.3.1 Let $X_1, \ldots, X_n$ be a random sample from a population with an unknown parameter θ. Then, a function $\varphi(X_1, \ldots, X_n)$ is called a *pivotal quantity* if it possesses the following properties:

(a) It is a function of sample values and some parameters, including the unknown parameter θ.
(b) Among all the parameters it contains, θ is the only unknown parameter.
(c) The probability distribution of $\varphi(X_1, \ldots, X_n)$ does not depend on the unknown parameter θ.

For example, let $X_1, \ldots, X_n$ be a random sample from a normal population with an unknown mean μ and known variance σ^2. Then, it can easily be verified that the random variable $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a pivotal quantity for μ. If μ and σ^2 are both unknown, then S^2/σ^2 is a pivotal quantity for σ^2.

Again dealing with the case that σ^2 is known and if $X_1, \ldots, X_n$ is a random sample from the normal distribution $N(\mu, \sigma^2)$, then from our discussion in chapter 7, it follows that the pivotal quantity $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has the standard normal distribution $N(0, 1)$. This means that we can write

$$P\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1 - \alpha \tag{8.3.1}$$

Solving the set of inequalities inside the parentheses, we can rewrite Equation (8.3.1) as

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{8.3.2}$$

Equation (8.3.2) essentially states that the random interval

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

contains μ with probability $(1 - \alpha)$. If σ is known, then the endpoints of the interval are known from information available from the sample and from Table A.4. We therefore state the following:

The interval estimator for the normal population mean with confidence coefficient $(1 - \alpha)$ when the population standard deviation σ is known is

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{8.3.3}$$

We sometimes say that Equation (8.3.3) is a $100(1 - \alpha)\%$ confidence interval for μ. The endpoints of the interval (8.3.3) are sometimes referred to as confidence limits, where

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

are called the *lower confidence limit* (LCL) and the *upper confidence limit* (UCL), respectively. Note that the confidence interval in Equation (8.3.3) can also be written as $\bar{X} \pm E$, where E is the margin of error defined in Equation (8.2.6). In words, Equations (8.3.2) and (8.3.3) state that if we were to observe a sample of n observations over and over again, and on each occasion, construct the confidence interval (8.3.3), then $100(1-\alpha)\%$ of these intervals (see Figure 8.3.1(a)) would contain the unknown population mean μ. This interpretation of confidence intervals is called the ***statistical interpretation***. Another interpretation of confidence intervals, called the ***practical interpretation***, is that we are $100(1-\alpha)\%$ confident that the confidence interval (8.3.3), obtained by using a single sample of size n, contains the unknown population mean μ (see Figure 8.3.1b).

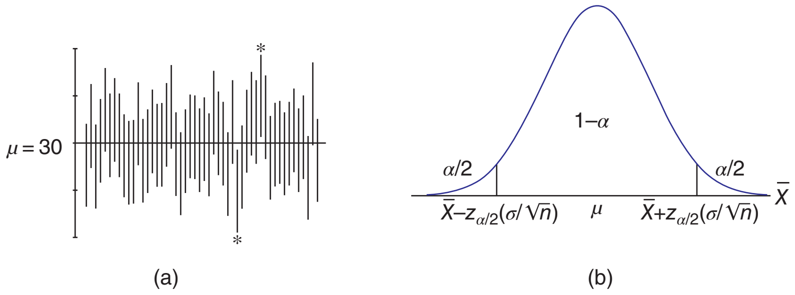


Figure 8.3.1 (a) Statistical and (b) practical interpretation of a confidence interval for the population mean μ.

Figure 8.3.1a shows 50 confidence intervals with confidence coefficient 95%, which we simulated by taking 50 random samples, each of size 25 from a normal population with mean 30 and variance 9. Each sample has a different sample mean $\bar{X}$, resulting in 50 confidence intervals centered at different points. Clearly, 48 of 50 intervals contain the true value of $\mu = 30$ (starred intervals do not contain $\mu = 30$), which means that 96% or just over 95% of the intervals, as expected, contain the true value of $\mu = 30$.

Figure 8.3.1b shows the practical interpretation of the confidence interval. That is, we are $100(1-\alpha)\%$ confident that the confidence interval obtained by using a single sample contains the unknown population mean μ. We formally state the above result as follows:

Theorem 8.3.1 *If $\bar{X}$ is the average of a sample of size n from a normal distribution $N(\mu, \sigma^2)$, where σ is known, then $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, or, more briefly, $\left(\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is a $100(1-\alpha)\%$ confidence interval for μ.*

Note that if the population from which the sample is drawn is not quite normal, but σ^2 is known and the sample size is large ($n \geq 30$), then $\left(\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is approximately a $100(1-\alpha)\%$ confidence interval for μ, since by the central limit theorem $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has, approximately, the standard normal distribution $N(0, 1)$, for large n.

Example 8.3.1 (Confidence interval for the mean using normal data with known variance) *A random sample of size 4 is taken from a population having the $N(\mu, 0.09)$ distribution. The observations were 12.6, 13.4, 12.8, and 13.2. Find a 95% confidence interval for μ.*

Solution: To find a 95% confidence interval for μ, we first note that

$$1-\alpha=0.95, \quad \alpha/2=0.025, \quad z_{\alpha/2}=z_{0.025}=1.96, \sigma=\sqrt{0.09}=0.3,$$
$$\frac{\sigma}{\sqrt{n}}=\frac{0.3}{\sqrt{4}}=0.15, \quad \bar{X}=\frac{52.0}{4}=13.0$$

Hence, the 95% confidence interval for μ is

$$(13.00 \pm 1.96(0.15))=(12.71,13.29)$$

The practical interpretation: We are 95% confident that the unknown population mean μ lies in the interval $(12.71, 13.29)$.

8.3.2 σ^2 Unknown

We turn now to the case where the sample is from a $N(\mu,\sigma^2)$ population where both μ and σ^2 are unknown and where we wish to estimate μ. Now, if $X_1, \ldots, X_n$ are n independent observations from the $N(\mu,\sigma^2)$ population, we know from Theorem 6.4.3 that $\bar{X}$ is distributed as $N(\mu,\sigma^2/n)$. Thus,

$$E(\bar{X})=\mu \tag{8.3.4}$$

That is, $\bar{X}$ is an unbiased point estimator of μ, whether σ^2 is known or not known. Now, of course, we cannot use Equation (8.3.3) as a confidence interval for μ when σ^2 is unknown. In practice, what we would like to do is to replace the unknown σ^2 by S^2 in Equation (8.3.1). However, if we do, we obtain $(\bar{X}-\mu)(S/\sqrt{n})$, a new random variable designated by T, and called the Student t variable. The distribution of T was first investigated by W.S. Gosset, writing under the pseudonym (Student, 1908). This early work was later verified by Fisher (1925). In Chapter 7 (see Theorem 7.3.7), we noted that the *pivotal quantity* $(\bar{X}-\mu)/(S/\sqrt{n})$ for μ is distributed as Student t with $n-1$ degrees of freedom. That is, if $(X_1, \ldots, X_n)$ is a random sample of n observations from $N(\mu,\sigma^2)$ and if

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i, \quad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

then, the random variable

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{8.3.5}$$

is a pivotal quantity and has a Student t-distribution with $n-1$ degrees of freedom. As discussed in Chapter 7, the t-distribution is symmetric around zero so that

$$P(-t_{n-1,\alpha/2} \le t_{n-1} \le t_{n-1,\alpha/2})=1-\alpha \tag{8.3.6}$$

where $P(t_{n-1} > t_{n-1,\alpha/2})=\alpha/2$. Using Equation (8.3.6), we may make the following statement:

$$P\left(-t_{n-1,\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}\right)=1-\alpha \tag{8.3.7}$$

which can be rewritten as

$$P\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha \tag{8.3.8}$$

Thus, if σ is unknown, we can make the following statement:

Theorem 8.3.2 *If $\bar{X}$ and S^2 are obtained from a random sample of size n from a normal distribution $N(\mu, \sigma^2)$, where both μ and σ are unknown, then $\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$, or, more briefly, $\left(\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$ is a $100(1-\alpha)\%$ confidence interval for μ.*

If the population from which the sample is drawn has an unknown mean and unknown variance but is not quite normal, $\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$ is an approximate $100(1-\alpha)\%$ confidence interval. This interval is often good enough for all practical purposes. We sometimes describe this by saying that the distribution of Equation (8.3.5) is robust, that is, not very sensitive to departures from the assumption that the distribution of X is normal.

We remind the reader that the use of Student's t-distribution to find a confidence interval for the population mean is applicable when the following conditions hold:

1. *The population is normal or approximately normal.*
2. *The sample size is small* ($n < 30$).
3. *The population variance is unknown.*

Note that some statisticians may prefer to use the t-distribution even if the sample size is larger than 30 because the t-values for any degrees of freedom are readily available using common statistical packages, such as MINITAB, R, Excel, or JMP.

Example 8.3.2 (Confidence interval for the mean using normal data with unknown variance) *Four determinations of the percentage of methanol in a certain solution yield $\bar{X} = 8.34\%, S = 0.03\%$. Assuming (approximate) normality of the population of determinations, find a 95% confidence interval for μ.*

Solution: To find a 95% confidence interval for μ, we note that $\bar{X} = 8.34, S/\sqrt{n} = 0.03/\sqrt{4} = 0.015$, and $1-\alpha = 0.95$, $\alpha/2 = 0.025$, $df = n-1 = 4-1 = 3$, $t_{3,0.025} = 3.182$.

Hence, a 95% confidence interval for μ is

$$(8.34 \pm 3.182(0.015)) = (8.292, 8.388)$$

(The value $t_{3,0.025} = 3.182$ is found using Table A.5.)

Interpretation: With 95% confidence, the average percentage of methanol in the solution is between 8.292% and 8.388%.

8.3.3 Sample Size Is Large

As is well known, the Student t-distribution approaches the $N(0, 1)$ distribution as the degrees of freedom tend to infinity. Thus, if n is large, we have, to a good approximation, that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim Z \tag{8.3.9}$$

where $Z \sim N(0, 1)$. Hence, if n is large, we may replace $t_{n-1,\alpha/2}$ by $z_{\alpha/2}$ in the statement of Theorem 8.3.2 and obtain an approximate $100(1 - \alpha)\%$ confidence interval for μ, namely

$$\left(\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \tag{8.3.10}$$

As a matter of fact, inspection of Tables A.4 and A.5, the tables of the standard normal and the t-distribution, respectively, shows that z_α and $t_{n,\alpha}$ are in "reasonably good agreement" for $n \geq 30$, and hence, we use Equation (8.3.10) for $n > 30$. This case is usually referred to as a *large-sample case.* ***Now with the availability of statistical packages, it is very easy to find the* t*-value for any degrees of freedom, one may continue using the* t*-distribution instead of switching to z even when*** $n > 30$.

Example 8.3.3 (Confidence interval for μ when using a large sample) *A manufacturing engineer decided to check the efficiency of a new technician hired by the company. She records the time taken by the technician to complete 100 randomly selected jobs and found that in this sample of 100, the average time taken per job was 10 hours with a standard deviation of two hours. Find a 95% confidence interval for μ, the average time taken by a technician to complete one job.*

Solution: In this example we do not know σ, but we are given that

$$\bar{X} = 10, \quad \text{and} \quad S = 2$$

Moreover, the sample size $n = 100$ is large. Thus, using the confidence interval $[\hat{\mu}_l, \hat{\mu}_u]$ given by Equation (8.3.10), we have

$$\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} = 10 - 1.96\frac{2}{\sqrt{100}} = 9.608,$$

and

$$\bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}} = 10 + 1.96\frac{2}{\sqrt{100}} = 10.392$$

Thus, a 95% confidence interval for the average time μ taken by a technician to complete one job is $(9.608, 10.392)$ hours. Note the confidence interval obtained here is usually called a *two-sided confidence interval.*

Interpretation: We are 95% confident that the average time the newly hired technician would take to complete the job is between 9.608 and 10.392 hours.

Note that the width of a confidence interval, which is defined as $\hat{\mu}_u - \hat{\mu}_l$, increases or decreases as the sample size decreases or increases, provided that the confidence coefficient remains the same.

Example 8.3.4 (Using MINITAB and R to find confidence interval for μ) *Consider the following data from a population with an unknown mean μ and unknown standard deviation σ:*

23 25 20 16 19 35 42 25 28 29 36 26 27 35 41 30

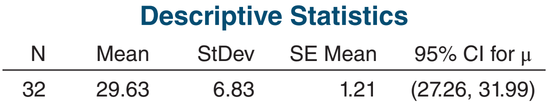
20 24 29 26 37 38 24 26 34 36 38 39 32 33 25 30

Use MINITAB and R to find a 95% confidence interval for the mean μ.

MINITAB

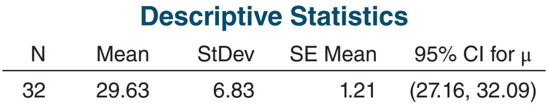
To find a 95% confidence interval for the mean μ using MINITAB, we proceed as follows:

1. Enter the data in column C1. If the summary statistics are given, then skip this step.
2. Since in this example the population standard deviation is not known, calculate the sample standard deviation of these data using one of the MINITAB procedures discussed earlier in Chapter 2. We will find $S = 6.82855$.
3. From the Menu bar select **Stat** > **Basic Statistics** > **1-Sample Z**. This prompts a dialog box **One-Sample Z for the Mean** to appear on the screen. Then form the pull down menu select **One or more samples, each in a column**. Enter C1 in the box below **One or more samples, each in a column**, and the value of standard deviation in a box next to **Standard deviation**. Note that since the sample size is greater than 30, we select the command **1-Sample Z** instead of **1-Sample t**.
4. Check **options**, which prompts another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level**, and under **alternative hypothesis** option, select **Mean $\neq$ hypothesized mean** by using the down arrow. In each dialog box, click **OK**. The MINITAB output will show up in the Session window as given below:



```
                Descriptive Statistics
 N      Mean     StDev   SE Mean    95% CI for μ
 32     29.63     6.83      1.21    (27.26, 31.99)
```

Note that if we had selected the command **1-Sample t**, then the output would be



```
                Descriptive Statistics
 N      Mean     StDev   SE Mean    95% CI for μ
 32     29.63     6.83      1.21    (27.16, 32.09)
```

Since the sample size is large, the two confidence intervals are almost the same. However, note that the confidence interval using the t-distribution is slightly larger, which is always the case.

USING R

The built in R function ‘z.test()’ in library ‘BSDA’ can be used to obtain the required z-interval. For the information provided in Example 8.3.4, a 95% confidence interval for the mean μ can be calculated by running the following in the R Console window.

```
install.packages("BSDA")
library(BSDA)


#Assign data
x = c(23,25,20,16,19,35,42,25,28,29,36,26,27,35,41,30,20,24,29,26,37,
         38,24,26,34,36,38,39,32,33,25,30)
z.test(x, alternative = "two.sided", mu = 0, sigma.x = sd(x),conf.level = 0.95)


#In case if we wanted to obtain the t-interval, use function
't.test()' in library 'stats' t.test(x, alternative = "two.sided",
mu = 0, sigma.x = sd(x), conf.level = 0.95)
```

These confidence intervals, of course, are identical (after rounding) to those produced by MINITAB.

One-Sided Confidence Interval

Note that the confidence interval in Equation (8.3.3) is symmetrical about the mean $\bar{X}$. This is because, while selecting the value of z, we divided α into two equal parts such that one half of α is used up in the lower tail and the other one half is used up in the upper tail. Technically speaking, we can divide α into two parts as we wish. That is, we may take, for example, one-third of α under one tail and the remaining two-thirds under the other tail. But, traditionally, we always divide α into two equal parts unless we have very strong reasons to do otherwise. Moreover, for a symmetric distribution, dividing α into two equal parts yields a slightly smaller confidence interval, which is one of the desirable properties of a confidence interval (the smaller the better). For one-sided confidence interval, we use all of α under one tail only. Thus, the two one-sided confidence intervals, commonly known as *upper one-sided* and *lower one-sided confidence intervals*, are given by

Upper one-sided and lower one-sided confidence intervals, respectively, for population mean with confidence coefficient $(1-\alpha)$ *when* σ *is known:*

$$(-\infty,\ \hat{\mu}_u) = (-\infty,\ \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}) \tag{8.3.11}$$

$$(\hat{\mu}_l,\ \infty) = (\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \infty) \tag{8.3.12}$$

Note that $\hat{\mu}_u$ and $\hat{\mu}_l$ are the upper and lower limits of a two-sided confidence interval with confidence coefficient $(1-2\alpha)$. Now, when the sample size is small and the sample is taken from a normal population with unknown variance, the *lower* and *upper one-sided* confidence intervals, respectively, are as follows:

Upper one-sided and lower one-sided confidence intervals, respectively, for population mean with confidence coefficient $(1-\alpha)$ *when* σ *is unknown, are respectively:*

$$\left(-\infty,\ \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \tag{8.3.13}$$

$$\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \infty\right) \tag{8.3.14}$$

The statistical and practical interpretations for one-sided confidence intervals are the same as for two-sided confidence intervals.

Example 8.3.5 (Finding one-sided confidence intervals) *Reconsider Example 8.3.3. Find 95%* lower *and* upper one-sided *confidence intervals for* μ*.*

Solution: In this example, $n = 100$, a large sample size, so we proceed as follows: $1-\alpha = 0.95, \alpha = 0.05$. Further, we know that $n-1 = 99$ and $t_{99;0.05} \approx Z_{0.05} = 1.645$ Then, from the results of Equations (8.3.13) and (8.3.14), the upper and lower one-sided confidence intervals for μ are, respectively, $(-\infty,\ \hat{\mu}_u)$ and $(\hat{\mu}_l,\ \infty)$ where in the present case $\hat{\mu}_l$ and $\hat{\mu}_u$ are

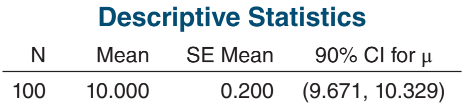
$$\hat{\mu}_l = \bar{X} - z_\alpha\frac{S}{\sqrt{n}} = 10 - 1.645\frac{2}{\sqrt{100}} = 9.671$$
$$\hat{\mu}_u = \bar{X} + z_\alpha\frac{S}{\sqrt{n}} = 10 + 1.645\frac{2}{\sqrt{100}} = 10.329$$

Thus, 95% one-sided upper and one-sided lower large sample confidence intervals for the population mean μ are $(-\infty, 10.329)$ and $(9.671, \infty)$, respectively.

Interpretation: We are 95% confident that the newly hired technician would take a maximum average time of 10.329 hours and a minimum average time of 9.671 hours to complete the job.

MINITAB

To find 95% one-sided confidence intervals for the population mean μ using MINITAB, we first find a 90% confidence interval using MINITAB. The MINITAB output for the 90% confidence interval is as follows (for Minitab instructions, see Example 8.3.4):



```
        Descriptive Statistics
  N     Mean    SE Mean    90% CI for μ
100   10.000      0.200    (9.671, 10.329)
```

Thus, 95% one-sided confidence intervals for the population mean μ are $(9.671, \infty)$ and $(-\infty, 10.329)$.

Alternatively, one can construct the one-sided upper/lower confidence interval by entering in the options dialog box 95% in the box next to **Confidence coefficient** and selecting less than/greater than under the **alternative** option.

USING R

The built in R function ‘z.test()’ in library ‘BSDA’ can be used to obtain the required z-interval. For the information provided in Example 8.3.3, a 95% confidence interval for the mean μ can be calculated by running the following in the R Console window.

```
install.packages(“BSDA”)
library(BSDA)


#Just like in MINITAB use confidence coefficient = 0.90
zsum.test(mean.x=10, sigma.x = 2, n.x = 100, alternative = “two.sided”,
mu = 0, conf.level = 0.90)
```

This confidence interval, of course, is identical (after rounding) to that produced by MINITAB.

Example 8.3.6 (Lower and upper confidence intervals for μ) *A random sample of size 25 of a certain kind of light bulb yielded an average lifetime of 1875 hours with a standard deviation of 100 hours. From past experience, it is known that the lifetime of this kind of bulb is normally distributed with mean μ and standard deviation σ. Find a 99% confidence interval for the population mean μ. Find 99% lower and upper one-sided confidence intervals for the population mean μ.*

Solution: From the given information, we find that in this particular example, the sample size is small, the population standard deviation σ is unknown, and the lifetime of the kind of bulb under study is normally distributed. Thus, in this example, we can use the Student t-distribution to find confidence intervals for the population mean μ. The summary statistics are

$$n = 25,\ \bar{X} = 1875,\ \text{and } S = 100,\ 1 - \alpha = 0.99,\ \alpha/2 = 0.005$$

Using the small-sample two-sided confidence interval $(\hat{\mu}_l, \hat{\mu}_u)$, we find that

$$\hat{\mu}_l = \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 1875 - t_{24,.005}\frac{100}{\sqrt{25}} = 1875 - 2.797 \times 20 = 1819.06$$
$$\hat{\mu}_u = \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 1875 + t_{24,.005}\frac{100}{\sqrt{25}} = 1875 + 2.797 \times 20 = 1930.94$$

Thus, a small-sample two-sided 99% confidence interval for μ is $(1819.06, 1930.94)$. The lower and upper one-sided 99% confidence limits are

$$\hat{\mu}_l = \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 1875 - t_{24,.01}\frac{100}{\sqrt{25}} = 1875 - 2.492 \times 20 = 1825.16$$

$$\hat{\mu}_u = \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = 1875 + t_{24,.01}\frac{100}{\sqrt{25}} = 1875 + 2.492 \times 20 = 1924.84$$

Thus, the 99% lower and upper one-sided small-sample confidence intervals for the population mean μ are $(1825.16, \infty)$ and $(0, 1924.84)$. Note that in the upper one-sided confidence interval, the lower limit is zero instead of being $-\infty$, since the lifetime of bulbs cannot be negative.

PRACTICE PROBLEMS FOR SECTION 8.3

1. The following data give the ages of 36 randomly selected family caregivers of older parents in the United States:

55	53	47	47	49	43	47	40	48	41	44	51	48	43	50	49	47	42
42	47	47	49	46	46	43	41	45	51	44	48	43	50	53	44	49	53

Assuming normality,

(a) Determine a 95% confidence interval for the mean ages of all US caregivers.

(b) Determine a one-sided lower and one-sided upper 95% confidence interval for the mean ages of all US caregivers.

2. Suppose that in Problem 1 only 25 of 36 randomly selected family caregivers responded, so we have the following data:

55	53	47	47	49	43	47	40	48	41	44	51	48
43	50	49	47	42	42	47	47	49	46	46	43	

Assuming that these data come from a population that has a normal distribution,

(a) Determine a 99% confidence interval for the mean ages of all US caregivers.

(b) Determine a one-sided lower and one-sided upper 99% confidence interval for the mean ages of all US caregivers.

3. An insurance company is interested in determining the average postoperative length of stay (in days) in hospitals for all patients who have bypass surgery. The following data give length of stay of 50 randomly selected patients who had bypass surgery:

6	10	10	9	9	12	7	12	7	8	10	7	8	8	10	12	10
7	7	10	8	8	10	7	6	12	9	7	8	6	6	10	6	7
7	10	8	12	8	10	7	7	10	11	11	8	6	7	8	11	

(a) Determine a 90% confidence interval for the mean postoperative length of stay (in days) in hospitals for all patients who have bypass surgery.

(b) Determine a one-sided lower and one-sided upper 90% confidence interval for the mean postoperative length of stay (in days) in hospitals for all patients who have bypass surgery.

4. A study was undertaken to see if the length of slide pins used in the front disc brake assembly met with specifications. To this end, measurements of the lengths of 16 slide pins, selected at random, were made. The average value of 16 lengths was 3.15, with a sample standard deviation of 0.2. Assuming that the measurements are normally distributed, construct a 95% confidence interval for the mean length of the slide pins.
5. The following data give the drying time (in hours) for 10 randomly selected concrete slabs:

9.06	9.17	9.11	8.16	9.10	9.98	8.89	9.02	9.32	8.12

   Assuming that drying times are normally distributed, determine a 95% confidence interval for the mean drying time for the slabs.
6. The weights of a random sample of 49 university male first-year students yielded a mean of 165 pounds and a standard deviation of 6.5 pounds. Determine a 90% confidence interval for the mean weight of all university male first-year students.
7. It is believed that drinking has some bad effects on the human reproduction system. To study this, some evaluations of placenta tissue of 16 randomly selected drinking mothers were made that yielded the following values (recorded to the nearest whole number):

18	17	22	21	15	21	22	22	14	20	14	16	13	22	20	19

   Assuming that these evaluations are normally distributed, determine a 99% confidence interval for the mean value for drinking mothers.
8. A hotel facility management company is interested in determining the average temperature during July at the location of one of their hotels. The temperatures of 49 randomly selected days in July during the past five years were as follows:

95	84	87	81	84	89	80	83	82	90	82	87	90	81	83	85	94
92	92	95	95	80	88	87	85	95	80	81	81	93	87	92	94	83
80	81	95	93	82	82	83	85	80	95	84	82	84	81	88		

   Determine a 99% confidence interval for the mean temperature for month of July at the place where the hotel is located.
9. Refer to Problem 8. Determine one-sided lower and one-sided upper 95% confidence intervals for the mean temperature for July where the hotel is located.
10. A sample of 25 bulbs is taken from a large lot of 40-watt bulbs, and the average of the sample bulb lives is 1410 hours. Assuming normality of bulb lives and that the standard deviation of bulb lives in the mass-production process involved is 200 hours, find a 95% confidence interval for the mean life of the bulbs in the lot.

11. A certain type of electronic condenser is manufactured by the ECA company, and over a large number of years, the lifetimes of the parts are found to be normally distributed with standard deviation $\sigma = 225$ h. A random sample of 30 of these condensers yielded an average lifetime of 1407.5 hours. Find a 99% confidence interval for μ, the mean lifetime of ECA condensers. What can you say about the statement "$\mu = 1400$"?

8.4 INTERVAL ESTIMATORS FOR THE DIFFERENCE OF MEANS OF TWO NORMAL POPULATIONS

8.4.1 Variances Are Known

Suppose that $\bar{X}_1$ and S_1^2 are the average and variance of a sample of size n_1 from the normal distribution $N(\mu_1, \sigma_1^2)$ and that $\bar{X}_2$ and S_2^2 are the average and variance of a sample of size n_2 from a normal distribution $N(\mu_2, \sigma_2^2)$. Consider the difference of the population means μ_1 and μ_2, say $\delta = \mu_1 - \mu_2$. The unbiased point estimator of δ is, of course, $\bar{X}_1 - \bar{X}_2$, since $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 = \delta$. Furthermore, from Theorem 7.3.8, we have that $\bar{X}_1 - \bar{X}_2$ is distributed by the $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ distribution. If σ_1^2 and σ_2^2 are known, then we may state that

$$P\left(-z_{\alpha/2} \le \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right) = 1 - \alpha \tag{8.4.1}$$

or equivalently,

$$P\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha \tag{8.4.2}$$

Hence, we have the following:

Theorem 8.4.1 *If $\bar{X}_1$ is the sample average of a random sample of size n_1 from a population having distribution $N(\mu_1, \sigma_1^2)$, and $\bar{X}_2$ is the sample average of an independent random sample of size n_2 from $N(\mu_2, \sigma_2^2)$, and if σ_1^2 and σ_2^2 are known, then*

$$\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{8.4.2a}$$

is a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

Note that when the two populations are normal with known variances, then the confidence interval for the difference of the two population means, shown in Equation (8.4.2a) is valid regardless of the sample sizes.

Example 8.4.1 (Constructing confidence interval for $\mu_1 - \mu_2$ with known variances) *A sample of size 10 from $N(\mu_1, 25)$ yields a sample average of $\bar{X}_1 = 19.8$, while an independent sample of size 12 from $N(\mu_2, 36)$ yields a sample average of $\bar{X}_2 = 24$. Find a 90% confidence interval for $\mu_1 - \mu_2$.*

Solution: In this example, the two populations are normal with known variances, and we have

$$1 - \alpha = 0.90,\ \alpha/2 = 0.05,\ z_{\alpha/2} = 1.645$$

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{25}{10} + \frac{36}{12}} = \sqrt{5.5} = 2.345$$

Hence, using Equation (8.4.2a), we obtain a 90% confidence interval for $\mu_1 - \mu_2$, which is

$$((19.8 - 24.0) \pm 1.645 \times 2.345) = (-9.06,\ -1.34)$$

Interpretation: We are 90% confident that the difference between the two population means $(\mu_1 - \mu_2)$ is between -9.06 and -1.34.

A word of caution: the 90% confidence interval for $\mu_2 - \mu_1$ is $(1.34, 9.06)$; that is, the sign of the confidence limits changes.

8.4.2 Variances Are Unknown

If n_1 and n_2 are large (both n_1 and n_2 are strictly greater than 30), and if σ_1^2 and σ_2^2 are unknown and we *cannot assume* $\sigma_1^2 = \sigma_2^2$, then

Confidence interval for $\mu_1 - \mu_2$ with confidence coefficient $(1 - \alpha)$ when σ_1^2 and σ_2^2 are unknown and we cannot assume $\sigma_1^2 = \sigma_2^2$ for large samples:

$$\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) \tag{8.4.3}$$

Here, Equation (8.4.3) is an approximate $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$. Further, this statement holds sufficiently well for most practical purposes, even if the two populations being sampled are fairly nonnormal, by virtue of the central limit theorem, since n_1 and n_2 are both large. If we can assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ^2 is unknown, it is usual to combine the two separate estimators S_1^2 (based on $\nu_1 = n_1 - 1$ degrees of freedom) and S_2^2 (based on $\nu_2 = n_2 - 1$ degrees of freedom) into a single estimator S_p^2 of the common variance σ^2. Whatever n_1 or n_2 may be, the pooled estimate S_p^2 is given by the weighted average of the individual estimators, the weights being the associated degrees of freedom, that is,

$$S_p^2 = \frac{\nu_1 S_1^2 + \nu_2 S_2^2}{\nu_1 + \nu_2} \tag{8.4.4}$$

Note that

$$E(S_p^2) = \frac{1}{\nu_1 + \nu_2} E(\nu_1 S_1^2 + \nu_2 S_2^2) = \frac{1}{\nu_1 + \nu_2} E(\nu_1 \sigma_1^2 + \nu_2 \sigma_2^2) = \sigma^2$$

It follows from Theorems 7.3.3 and 7.3.5 that the quantity

$$(n_1 + n_2 - 2)\frac{S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_1-1}^2 + \chi_{n_2-1}^2 \tag{8.4.5}$$

where the $\chi_{n_1-1}^2$ and $\chi_{n_1-1}^2$ are independent chi-squared random variables, so that $(n_1 + n_2 - 2)S_p^2/\sigma^2$ has a chi-square distribution with $n_1 + n_2 - 2$ degrees of freedom. Here, we recognize that S_p^2 is an unbiased point estimator of σ^2, sometimes called the *pooled estimator* of σ^2. Furthermore, we know that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8.4.6}$$

has the $N(0, 1)$ distribution. Also, the quantities in Equations (8.4.5) and (8.4.6) are independent. Thus, from Theorem 7.3.9, it follows that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{8.4.7}$$

has the Student t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Therefore, we can say that

$$P\left(-t_{n_1+n_2-2;\alpha/2} \le \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le t_{n_1+n_2-2;\alpha/2}\right) = 1 - \alpha \tag{8.4.8}$$

or, alternatively,

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2)\right.$$

$$\left. + \, t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha \tag{8.4.9}$$

We summarize the preceding result as follows:

Theorem 8.4.2 *If in Theorem 8.4.1 it is assumed that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ^2 is unknown, then*

$$\left((\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \tag{8.4.10}$$

is a $100(1-\alpha)\%$ *confidence interval for* $\mu_1 - \mu_2$.

Example 8.4.2 (Constructing a confidence interval for $\mu_1 - \mu_2$ with unknown but equal variances) *A sample of $n_1 = 5$ light bulbs of type A gives an average length of life of $\bar{X}_1 = 1000$ hours with a standard deviation of $S_1 = 28$ hours. A sample of $n_2 = 7$ light bulbs of type B, gives $\bar{X}_2 = 980$ hours, and $S_2 = 32$ hours. We assume that the processes are normally distributed with variances σ_1^2 and σ_2^2 that are equal, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Find a 99% confidence interval for $\mu_1 - \mu_2 = \mu_A - \mu_B$.*

Solution: In this example, we have two normal populations with equal variances that are unknown. Sample sizes are small; therefore, we use the Student t-distribution for determining a 99% confidence interval for $\mu_1 - \mu_2$, Further, we note that

$$1 - \alpha = 0.99,\ \alpha/2 = 0.005,\ n_1 + n_2 - 2 = 10,\ t_{10,0.005} = 3.169$$

and that

$$S_p^2 = \frac{(5-1)28^2 + (7-1)32^2}{10} = 928 \text{ or } S_p = 30.46$$

Hence, the desired confidence interval for $\mu_1 - \mu_1$ is given by

$$\left((1000 - 980) \pm 3.169 \times 30.46 \times \sqrt{\frac{1}{5} + \frac{1}{7}} \right) = (20 \pm 56.5) = (-36.5, 76.6)$$

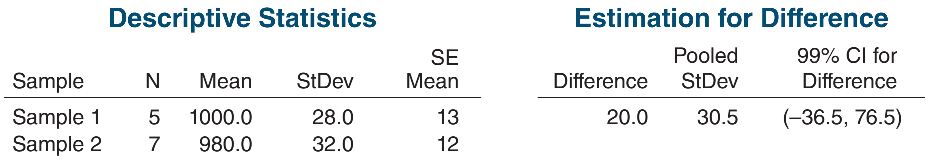
Interpretation: We are 99% confident that the difference between the two population means ($\mu_A - \mu_B$) is between 36.5 and 76.6.

MINITAB

To find a 99% confidence interval for the mean using MINITAB, we proceed as follows:

1. From the Menu bar, select **Stat** > **Basic Statistics** > **2-Sample t**. This prompts a dialog box **Two-Sample t for the Mean** to appear on the screen.
2. Select **Summarized data** from the pull down menu, and then enter the values of sample size, sample mean, and sample standard deviation in the respective boxes.
3. Check **Options**, which will prompt another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level** and select **Difference ≠ hypothesized difference** under **alternative** option. If the variances are equal,

then check the box next to **Assume equal variances**. Otherwise, do not check that box. Click **OK** on each of the two dialog boxes. The Minitab output will show up in the Session window as follows:



Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	5	1000.0	28.0	13
Sample 2	7	980.0	32.0	12

Estimation for Difference

Difference	Pooled StDev	99% CI for Difference
20.0	30.5	(–36.5, 76.5)

Note that this result matches the result obtained manually. Also, if in the second dialog box we check **Difference** < **hypothesized difference** or **Difference** > **hypothesized difference** instead of **Difference** ≠ **hypothesized difference**, we obtain upper and lower bounds for one-sided confidence intervals, respectively.

USING R

The built in R function 'tsum.test()' in library 'BSDA' can be used to conduct two-sample t-test. For the information provided in Example 8.4.2, the test can be conducted by running the following in the R Console window.

```
install.packages("BSDA")
library(BSDA)
tsum.test(mean.x = 1000, s.x = 28, n.x = 5, mean.y = 980, s.y = 32,
n.y = 7, alternative = "two.sided", mu = 0, var.equal = TRUE,conf.level = 0.99)


#R output
Standard Two-Sample t-Test
data: Summarized x and y
t = 1.1212, df = 10, p-value = 0.2884
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
-36.53146 76.53146
sample estimates:
mean of x mean of y
1000, 980
```

This confidence interval, of course, is identical (after rounding) to that produced manually.

Recall from Section 7.3 that for large degrees of freedom $m, t_m \cong z$. In the two-sample problem, if $(n_1 + n_2 - 2) \geq 60$, that is, if $n_1 + n_2 \geq 62$, then an approximate $100(1-\alpha)\%$ interval for $\mu_1 - \mu_2$ is obtained by using z instead of t. That is, we would use

$$\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \tag{8.4.11}$$

when $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

We now discuss a method available for the situation when $\sigma_1^2 \neq \sigma_2^2$. The procedure is as follows: compute m, the degrees of freedom from Satterthwaite's approximation for t given by

$$m = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} \tag{8.4.12}$$

An equivalent method to calculate the degrees of freedom is as follows: calculate

$$c = \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}$$

then m is such that

$$\frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}$$

Now it can be shown that an approximate confidence interval for $\mu_1 - \mu_2$ with confidence coefficient $(1-\alpha)$ when σ_1^2 and σ_2^2 are unknown and cannot assume $\sigma_1^2 = \sigma_2^2$ for small samples is:

$$\left((\bar{X}_1 - \bar{X}_2) - t_{m,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + t_{m,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) \tag{8.4.13}$$

Further, it can be shown that

$$\min(n_1 - 1, n_2 - 1) \leq m \leq n_1 + n_2 - 2 \tag{8.4.14}$$

Hence, if n_1 and n_2 are *both strictly greater* than 30, we use Equations (8.4.11) and (8.4.12) and obtain an m that is greater than or equal to 30. This in turn means that the probability point $t_{m,\alpha/2}$ used in the interval (8.4.13) is such that $t_{m,\alpha/2} \approx z_{\alpha/2}$, and (8.4.13) takes the form

$$\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) \tag{8.4.15}$$

so that Equation (8.4.15) is an appropriate $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ when $n_1 > 30$, $n_2 > 30$, and $\sigma_1^2 \neq \sigma_2^2$.

Example 8.4.3 (Constructing a confidence interval for $\mu_1 - \mu_2$ with variances unknown and unequal) *Twenty-five batch yields of a plastic produced using a catalyst (method 1) are to be compared with 25 batch yields of the plastic produced without the catalyst (method 2). The following results were obtained (coded units):*

$$\textit{Method 1}: n_1 = 25,\ \bar{X}_1 = 6.40,\ S_1^2 = 2.4264$$

$$\textit{Method 2}: n_2 = 25,\ \bar{X}_2 = 6.02,\ S_2^2 = 1.0176$$

Assuming normality of batch yields obtained using method i, *say* $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, *find a 95% confidence interval for* $(\mu_1 - \mu_2)$. *Assume that* $\sigma_1^2 \neq \sigma_2^2$.

Solution: Since it is given that $\sigma_1^2 \neq \sigma_2^2$, we proceed to find a 95% confidence interval for $\mu_1 - \mu_2$ using the interval (8.4.13). For this, we must first determine the value of m, which we expect from Equation (8.4.14) to lie between 24 and 48 inclusive. From Equation (8.4.12), we have

$$m = \frac{\left(\frac{2.4264}{25} + \frac{1.0176}{25}\right)^2}{\frac{\left(\frac{2.4264}{25}\right)^2}{24} + \frac{\left(\frac{1.0176}{25}\right)^2}{24}} = 41.1$$

We use $m = 41$. Now we have $1 - \alpha = 0.95$, $\alpha/2 = 0.025$. Thus from Table A.5 we have $t_{41,.025} \approx 2.0195$. Hence, a 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\left((6.40 - 6.02) \pm 2.0195 \times \sqrt{\frac{2.4264}{25} + \frac{1.0176}{25}}\right) = (0.38 \pm 0.75) = (-0.37, 1.13)$$

Interpretation: We are 95% confident that the difference between the two population means $(\mu_1 - \mu_2)$ is between -0.37 and 1.13.

We note that with confidence coefficient $1 - \alpha = 0.95$, the confidence interval contains 0. Therefore, the sample evidence supports the statement $\mu_1 - \mu_2 = 0$, that is, $\mu_1 = \mu_2$. (We return to this point in Chapter 9, where we use the confidence interval as an alternate method for testing a certain simple hypothesis.)

Example 8.4.4 (Yarn breaking strength test) *A sample of 61 strands of type I yarn, when subjected to breaking-strength tests, yielded a sample average of* $\bar{X}_I = 1400$ *psi with a sample standard deviation of* $S_I = 120$ *psi. A sample of 121 strands of type M yarn was also subjected to the same breaking strength tests and yielded a sample average of* $\bar{X}_M = 1250$ *psi and a sample standard deviation of* $S_M = 80$ *psi. Find a 95% confidence interval for* $\mu_I - \mu_M$, *assuming normality of breaking strengths of both types of yarn. Assume that* $\sigma_I^2 \neq \sigma_M^2$.

Solution: Since we cannot assume that $\sigma_I^2 = \sigma_M^2$, and both sample sizes are large, we proceed first to find a 95% confidence interval for $\mu_I - \mu_M$ using Equation (8.4.15). In fact, if we went through the calculations of Equation (8.4.12), the reader should verify that we would find that $m = 87$ and $t_{87,\alpha} \approx z_\alpha$. Hence the 95% confidence interval for $\mu_I - \mu_M$ is

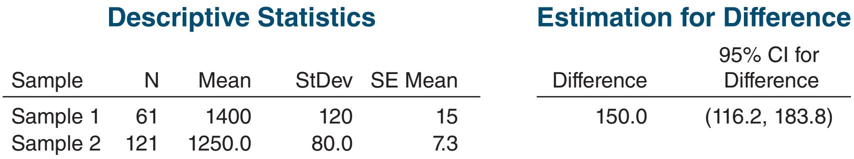
$$\left((1400 - 1250) \pm 1.96 \times \sqrt{\frac{120^2}{61} + \frac{80^2}{121}}\right) = (150 \pm 1.96 \times \sqrt{289}) = (116.8,\ 183.2)$$

This means that we are 95% confident that the difference between breaking strength of two types of yarn is between 116.8 and 183.2 psi.

In this example, the confidence interval with confidence coefficient $1 - \alpha = 0.95$ does not contain 0, and both confidence limits are positive. Therefore, the sample evidence supports the statement $\mu_I - \mu_M > 0$, that is, $\mu_I > \mu_M$. We now complete this example using MINITAB and R.

MINITAB

To find a 95% confidence interval for $\mu_I - \mu_M$ using MINITAB, we proceed in the same manner as in Example 8.4.2. The MINITAB output shows up in the Session window as:



```
Descriptive Statistics                                  Estimation for Difference

                                                                       95% CI for
Sample      N     Mean    StDev  SE Mean                Difference     Difference
Sample 1   61     1400      120       15                     150.0  (116.2, 183.8)
Sample 2  121   1250.0     80.0      7.3
```

Note that these confidence limits are slightly different from the limits we found previously for this example. This difference can be expected since we used z-values, whereas R and MINITAB use t-values. Further note that if instead of summary statistics we are given raw data, then in the **Two-Sample t for the Mean** dialog box, we would select the option data in one column or data in two columns depending on whether we have entered the data in one column or two columns.

USING R

As in Example 8.4.2, we can use the built in R function ‘tsum.test()’ in library ‘BSDA’ with specifying ‘var.equal = FALSE’.

```
install.packages("BSDA")
library(BSDA)
tsum.test(mean.x = 1400, s.x = 120, n.x = 61, mean.y = 1250,s.y = 80, n.y = 121,
alternative = "two.sided", mu = 0, var.equal = FALSE, conf.level = 0.95)
```

PRACTICE PROBLEMS FOR SECTION 8.4

1. A sample of size 10 from $N(\mu_1, 225)$ yields an average $\bar{X}_1 = 170.2$, while an independent sample of size 12 from $N(\mu_2, 256)$ yields a sample average $\bar{X}_2 = 176.7$. Find a 95% confidence interval for $\mu_1 - \mu_2$.
2. Suppose that random samples of size 25 are taken from two large lots of light bulbs, say lot A and lot B, and the average lives for the two samples are found to be

$$\bar{X}_A = 1580 \text{ hours}, \quad \bar{X}_B = 1425 \text{ hours}$$

   Assuming that the standard deviation of bulb life in each of the two lots is 200 hours, find 95% confidence limits for $\mu_A - \mu_B$.
3. A light bulb company tested ten light bulbs that contained filaments of type A and ten that contained filaments of type B. The following results were obtained for the length of life in hours of the 20 light bulbs (Steele and Torrie):

Filament A:	1293	1380	1614	1497	1340	1643	1466	1094	1270	1028
Filament B:	1061	1627	1065	1383	1092	1711	1021	1138	1017	1143

Assuming (approximate) normality and equal variances, find a 95% confidence interval for the difference between the mean life of bulbs with filament A and filament B. Does the sample evidence support the assumption that these means are equal?

4. Before and after tests are to be made on the breaking strengths (oz) of a certain type of yarn. Specifically, seven determinations of the breaking strength are made on test pieces of the yarn before a spinning machine is reset and five on test pieces after the machine is reset, with the following results:

Before:	22.7	25.7	20.7	26.7	21.2	19.2	22.7
After:	23.2	23.7	25.2	23.7	24.7		

Assuming that determinations made "after reset" and determinations made "before reset" are independently distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, find a 95% confidence interval for $\mu_1 - \mu_2$ without assuming equal variances.

5. A new catalyst is to be used in production of a plastic chemical. Twenty batches of chemical are produced, 10 batches with the new catalyst, 10 without. The results are as shown below:

Catalyst present:	7.2	7.3	7.4	7.5	7.8	7.2	7.5	8.4	7.2	7.5
Catalyst absent:	7.0	7.2	7.5	7.3	7.1	7.1	7.3	7.0	7.0	7.3

Assuming normality and that variances are not equal, find a 99% confidence interval for the difference of the means of yields of batches obtained when using the catalyst and when not using the catalyst. On the basis of this interval, do the data support the claim that use of the catalyst increases the average batch yield?

6. Reconsider Problem 5 of Section 8.3. Suppose now that the drying time is measured at two different temperature settings T_1 and T_2. The data obtained are as follows:

T_1:	9.93	9.66	9.11	9.45	9.02	9.98	9.17	9.27	9.63	9.46
T_2:	10.26	9.75	9.84	10.42	9.99	9.98	9.89	9.80	10.15	10.37

Assuming that the drying times for both temperature settings are normally distributed with equal variances, determine a 95% confidence interval for the difference of means of drying times when settings T_1 and T_2 are used.

7. Repeat Problem 6 above, now assuming that the two population variances are not equal.
8. The following data give the LDL cholesterol (commonly known as bad cholesterol) levels of two groups I and II of young female adults. Each member of group I followed a very strict exercise regimen, whereas group II members did not do any exercise.

Group I:	85	84	76	88	87	89	80	87	71	78
	74	80	89	79	87	75	83	84	71	70
Group II:	91	105	98	98	98	107	101	101	94	96
	103	109	105	103	95	97	95	91	104	107

   Suppose that μ_1 and μ_2 are the means of the two populations from which young female adults in groups I and II have been selected. Assuming that the two populations are normally distributed with equal variance, determine a 99% confidence interval for $\mu_1 - \mu_2$.
9. Suppose in Problem 8 that the variances of the two populations are known from past experience to be 16 and 25, respectively. Determine a 99% confidence interval for $\mu_1 - \mu_2$.
10. Two different brands of an all-purpose joint compound are used in residential construction and their drying times, in hours, are recorded. Sixteen specimens for each joint compound were selected. Recorded drying times are:

Brand I:	11.19	10.22	10.29	11.11	10.08	10.14	10.60	10.08
	11.28	11.98	11.22	11.97	10.47	10.79	11.98	10.03
Brand II:	12.10	13.91	13.32	13.58	12.04	12.00	13.05	13.70
	12.84	13.85	13.40	12.48	13.39	13.61	12.37	12.08

   Assuming that the drying times of the two brands are normally distributed with equal variances, determine a 95% confidence interval for $\mu_{\mathrm{I}} - \mu_{\mathrm{II}}$.

8.5 INTERVAL ESTIMATORS FOR THE VARIANCE OF A NORMAL POPULATION

We saw in Chapter 7 that if S^2 is the variance of a sample of size n from a normal distribution $N(\mu, \sigma^2)$, then $(n-1)S^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom. Note that as the degree of freedom changes, the shape of the chi-square distribution changes (see Figure 8.5.1).

We can now state, using the distributional result mentioned previously, that

$$P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) = 1 - \alpha \qquad (8.5.1)$$

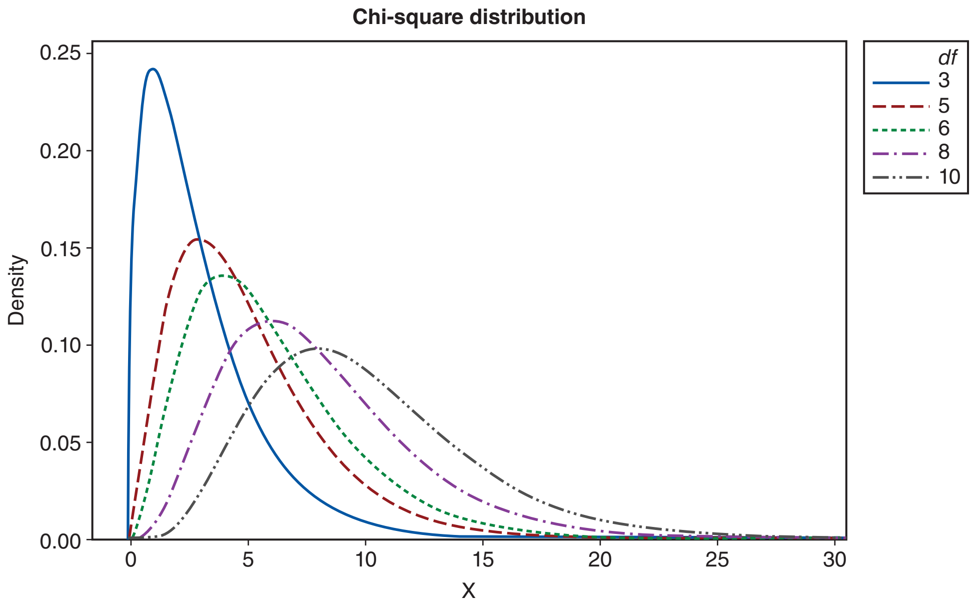


Figure 8.5.1 Chi-square distribution for various degrees of freedom.

or equivalently

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha \tag{8.5.2}$$

This enables us to state the following:

Theorem 8.5.1 *If S^2 is the sample variance of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$, then a $100(1-\alpha)\%$ confidence interval for σ^2 is given by*

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) \tag{8.5.3}$$

Referring to Equation (8.5.2), it should be noted that the confidence interval for σ can be obtained simply by taking the square root of the confidence limits of σ^2.

A $100(1-\alpha)\%$ confidence interval for σ is given by

$$\left(S\sqrt{\frac{(n-1)}{\chi^2_{n-1,\alpha/2}}}, S\sqrt{\frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}}}\right) \tag{8.5.4}$$

The confidence interval (8.5.4) contains σ if and only if the confidence interval (8.5.3) contains σ^2.

Now it turns out that the confidence interval for σ^2 (or σ) can be quite off the mark, as contrasted to that for μ, if the population distribution departs significantly from normality. Note that $100(1-\alpha)\%$ *one-sided lower* and *upper confidence intervals* for σ^2 are given by

One-sided lower and upper confidence intervals for σ^2 with confident coefficient $(1-\alpha)$

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}, \infty\right) \text{ and } \left(0, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}\right) \tag{8.5.5}$$

respectively. A one-sided lower and upper confidence interval for σ are found by taking the square roots of the corresponding limits in Equation (8.5.5).

Example 8.5.1 (Confidence interval for σ^2) *Refer to Example 8.3.2, and find a 95% confidence interval for σ^2.*

Solution: From Example 8.3.2, we have $n-1=3, S^2=0.0009, 1-\alpha=0.95, and$ $1-\alpha/2=0.975$. We then need $\chi^2_{3,0.025}=9.348$ and $\chi^2_{3,0.975}=0.216$, which can be found using Appendix A.6. This implies that a 95% confidence interval for σ^2 is given by

$$\left(\frac{3(0.0009)}{9.348}, \frac{3(0.0009)}{0.216}\right) = (0.00029, 0.0125)$$

Note that a 95% confidence interval for σ is $(0.017, 0.112)$.

Interpretation: We are 95% confident that population variance σ^2 falls between 0.00029 and 0.0125, and the population standard deviation σ falls between 0.017 and 0.112.

Example 8.5.2 (Lower and upper confidence intervals for σ^2) *Referring to Example 8.5.1, find 95% one-sided lower and upper confidence intervals for σ^2 and σ.*

Solution: As in Example 8.5.1, we have $n-1=3, S^2=0.0009, 1-\alpha=0.95,$ so $\chi^2_{3,0.05}=7.81$ and $\chi^2_{3,0.95}=0.35$. Hence, the 95% lower and upper one-sided confidence intervals $[\hat{\sigma}^2_l, \infty]$ and $[0, \hat{\sigma}^2_u]$ for σ^2 are such that

$$\hat{\sigma}^2_l = \left(\frac{3(0.0009}{7.81}\right) = 0.000346,\ \hat{\sigma}^2_u = \left(\frac{3(0.0009}{0.35}\right) = 0.0077$$

respectively. Now, by taking the square root of these confidence interval limits, it can be seen that the 95% lower and upper one-sided confidence intervals for σ are $(0.0186, \infty)$ and $(0, 0.0877)$, respectively.

Example 8.5.3 (Confidence interval for σ^2 and σ) *In a certain car-manufacturing company, the time taken by a worker to finish a paint job on a car is normally distributed with mean μ and variance σ^2. Fifteen randomly selected car paint jobs are assigned to that* *worker, and the time taken by the worker to finish each job is jotted down. These data yielded a sample standard deviation of $S = 2.5$ hours. Find a 95% two-sided confidence interval and one-sided lower and upper confidence intervals for the population variance σ^2 and the standard deviation σ.*

Solution: From the given information and using the chi-square distribution table, we have $S = 2.5$, $\alpha = 0.05$, $n - 1 = 14$, $\chi^2_{14,0.025} = 26.119$, $\chi^2_{14,0.975} = 5.629$ Hence, a 95% confidence interval for σ^2 is given by

$$\left(\frac{14(2.5)^2}{26.119}, \frac{14(2.5)^2}{5.629} \right) = (3.35, 15.54)$$

A 95% confidence interval for σ is found by taking the square root of the corresponding limits for σ^2 Thus, a 95% confidence interval for σ is (1.83, 3.94).

To find a one-sided confidence interval, note that the value of the χ^2 point changes, since the whole value of α falls under one tail only. For example, we have

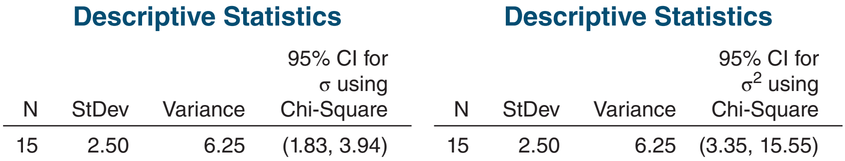
$$\hat{\sigma}_l^2 = \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}} = \frac{(15-1)(2.5)^2}{\chi^2_{14,0.05}} = \frac{87.5}{23.685} = 3.69$$
$$\hat{\sigma}_u^2 = \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}} = \frac{(15-1)(2.5)^2}{\chi^2_{14,0.95}} = \frac{87.5}{6.57} = 13.32$$

Therefore, one-sided lower and upper 95% confidence intervals for σ^2 are $(3.69, \infty)$ and (0, 13.32), respectively. The confidence intervals for the population standard deviation are found by taking the square root of the corresponding limits. That is, one-sided lower and upper 95% confidence intervals for the population standard deviation σ are $(1.92, \infty)$ and (0, 3.65), respectively. We now describe how to find these confidence interval using MINITAB and R.

MINITAB

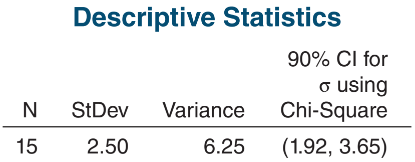
To find two-sided and one-sided confidence intervals for σ^2 using MINITAB, we proceed as follows. For part (a):

1. From the Menu bar, select **Stat** > **Basic Statistics** > **1-Variance**. This prompts a dialog box **One-Sample Variance** to appear on the screen.
2. Pull down the menu in this dialog box and select **sample standard deviation**; then make the appropriate entries in the boxes that follow.
3. Check **Options**, which prompts another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level** and under **Alternative hypothesis**, select **Standard deviation ≠ hypothesized standard deviation**. Click **OK** in each of the two dialog boxes. The MINITAB output shows up in the Session window as given below.



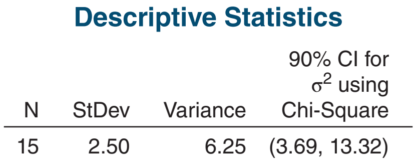
```
Descriptive Statistics                       Descriptive Statistics

                          95% CI for                                   95% CI for
                           σ using                                     σ² using
 N   StDev   Variance   Chi-Square          N   StDev   Variance     Chi-Square
----------------------------------         -----------------------------------
15    2.50       6.25   (1.83, 3.94)       15    2.50       6.25   (3.35, 15.55)
```

4. By repeating Step 2 by selecting **sample variance** from the pull down menu and making the appropriate entries in the boxes that follow, the confidence interval for variance can be obtained.
5. For part (b). To find 95% one-sided confidence intervals, we first find 90% two-sided confidence intervals for σ and σ^2, which turn out to be:



Descriptive Statistics

N	StDev	Variance	90% CI for σ using Chi-Square
15	2.50	6.25	(1.92, 3.65)



Descriptive Statistics

N	StDev	Variance	90% CI for σ² using Chi-Square
15	2.50	6.25	(3.69, 13.32)

Thus, the lower and upper 95% one-sided confidence intervals for the standard deviation are $(1.92, \infty)$ and (0, 3.65) and for the variance are (3.69, ∞) and (0, 13.32), respectively.

USING R

We can use ‘qchisq(p, df, ncp = 0, lower.tail = TRUE)’ function in R to get the chisquare quantiles in the confidence interval calculations as follows.

```
#Assign variables
alpha = 0.05; S = 2.5; n = 15


#To obtain the two-sided confidence interval for σ^2
CI = c((n-1)*S^2/qchisq(1-alpha/2, n-1), (n-1)*S^2/qchisq(alpha/2,n-1))
CI #R output
[1] 3.350058, 15.545258


#To obtain the two-sided confidence interval for σ
sqrt(CI) #R output
[1] 1.830316, 3.942748


#To obtain the one-sided confidence interval for σ^2
Lower.limit = (n-1)*S^2/qchisq(1-alpha, n-1)
Lower.limit #R output
[1] 3.694354
Upper.limit = (n-1)*S^2/qchisq(alpha, n-1)
Upper.limit #R output
[1] 13.31683


#To obtain the one-sided confidence interval for σ
sqrt(Lower.limit) #R output
[1] 1.92207
sqrt(Upper.limit) #R output
[1] 3.649224
```

As found earlier, the two-sided 95% confidence interval is (1.83, 3.94) for the standard deviation and (3.35, 15.55) for the variance. Also, the lower and upper 95% confidence intervals for the standard deviation are, respectively, (1.92, ∞), and (0, 3.65), while for the variance, these are (3.69, ∞), (0, 13.32).

8.6 INTERVAL ESTIMATOR FOR THE RATIO OF VARIANCES OF TWO NORMAL POPULATIONS

Suppose that S_1^2 and S_2^2 are sample variances of two independent random samples from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, where σ_1^2 and σ_2^2 are unknown, and we want to find an interval estimator for the ratio σ_1^2/σ_2^2. We would proceed as follows.

We know from Chapter 7 that if the two samples are independently drawn from the two normal distributions indicated previously, then the quantities

$$\frac{(n_1-1)S_1^2}{\sigma_1^2}, \ \frac{(n_2-1)S_2^2}{\sigma_2^2} \tag{8.6.1}$$

are independent random variables having chi-square distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively. By the definition of Snedecor's F-distribution, we see that the random variable

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{8.6.2}$$

is distributed by the F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

Thus, we may write

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \le \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \le F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha \tag{8.6.3}$$

or solving the inequalities, we write

$$P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) = 1-\alpha \tag{8.6.4}$$

This leads us to the following theorem:

Theorem 8.6.1 *If S_1^2 and S_2^2 are variances of independent random samples of size n_1 and n_2 from the normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, where μ_1, μ_2, σ_1^2, and σ_2^2 are unknown, then a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is given by*

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\alpha/2}}, \ \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) \tag{8.6.5}$$

Note that $100(1-\alpha)\%$ one-sided lower and upper confidence intervals for the ratio of two population variances σ_1^2/σ_2^2 are as given below by Equations (8.6.6a) and (8.6.6b), respectively.

One-sided $100(1-\alpha)\%$ *lower and upper confidence intervals for the ratio of two population variances* σ_1^2/σ_2^2:

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\alpha}},\ \infty\right) \tag{8.6.6a}$$

$$\left(0,\ \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\alpha}}\right) \tag{8.6.6b}$$

Furthermore, a $100(1-\alpha)\%$ confidence interval for the ratio of two population standard deviations σ_1/σ_2 is found by taking the square roots of the corresponding limits in Equations (8.6.5), (8.6.6a), and (8.6.6b). This yields the following:

$100(1-\alpha)\%$ *confidence interval for the ratio of the two population standard deviations* σ_1/σ_2:

$$\left(\frac{S_1}{S_2}\frac{1}{\sqrt{F_{n_1-1,n_2-1,\alpha/2}}},\ \frac{S_1}{S_2}\frac{1}{\sqrt{F_{n_1-1,n_2-1,1-\alpha/2}}}\right) \tag{8.6.7}$$

and a $100(1-\alpha)\%$ one-sided lower and upper one-sided confidence intervals for the ratio of two population standard deviations σ_1/σ_2 are given by Equations (8.6.8a) and (8.6.8b), respectively.

One-sided $100(1-\alpha)\%$ *lower and upper confidence intervals for the ratio of two population standard deviations* σ_1/σ_2:

$$\left(\frac{S_1}{S_2}\frac{1}{\sqrt{F_{n_1-1,n_2-1,\alpha}}},\ \infty\right) \tag{8.6.8a}$$

$$\left(0,\ \frac{S_1}{S_2}\frac{1}{\sqrt{F_{n_1-1,n_2-1,1-\alpha}}}\right) \tag{8.6.8b}$$

Note that the usual F tables do not provide values of $F_{m_1,m_2,1-\alpha}$ for $\alpha \leq 0.10$, which means $1-\alpha \geq 0.90$. Thus in these cases, to determine the lower percentage points of the F-distribution, we use the following relation:

$$F_{m_1,m_2,1-\alpha} = \frac{1}{F_{m_2,m_1,\alpha}} \tag{8.6.9}$$

Example 8.6.1 (Confidence intervals for the ratio of two population variances) *Two random samples of sizes 13 and 16 are selected from a group of patients with hypertension. The patients in the two samples are independently treated with drugs A and B. After a full course of treatments, these patients are evaluated. The data collected at the time of evaluation yielded sample standard deviations $S_A = S_1 = 6.5$ mmHg and $S_B = S_2 = 7.5$ mmHg. Assuming that the two sets of data come from independent normal populations with variances σ_1^2 and σ_2^2, respectively, determine a 95% two-sided confidence interval for σ_1^2/σ_2^2. Also determine the one-sided confidence intervals for σ_1^2/σ_2^2 and σ_1/σ_2.*

Solution: From the information given, we have $1-\alpha = 0.95, n_1 - 1 = 12, n_2 - 1 = 15$, $S_1 = 6.5$, $S_2 = 7.5$, $F_{12,15,0.025} = 2.9633$ and using Equation (8.6.9), $1/F_{12,15,0.975} = F_{15,12,0.025} = 3.1772$.

Now using Equation (8.6.5), we obtain a 95% confidence interval for σ_1^2/σ_2^2 as

$$\left(\frac{6.5^2}{7.5^2}\frac{1}{2.9633}, \frac{6.5^2}{7.5^2}(3.1772)\right) = (0.25347,\ 2.386)$$

and for the ratio of the standard deviations σ_1/σ_2 the two-sided confidence interval is found by taking square roots, that is, (0.5034, 1.5446).

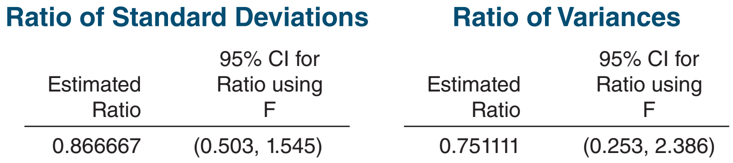
Now $F_{12,15,0.05} = 2.4753$ and $F_{15,12,0.05} = 2.6169$, so the 95% lower and upper one-sided confidence intervals for σ_1^2/σ_2^2 and σ_1/σ_2 are, as the reader should verify, given by $(0.3034, \infty)$, $(0, 1.9656)$ and $(0.5508, \infty)$, $(0, 1.4020)$, respectively.

The statistical and practical interpretation of all these confidence intervals is similar as for confidence intervals for the means. Further, the two-sided confidence intervals contain 1, which means that at the 5% level of significance, we can conclude that two variances are not significantly different from each other. There is more discussion on this aspect of our conclusion in Chapter 9 where we study the testing of hypotheses about two variances. The result above may be obtained by using MINITAB and R as follows:

MINITAB

To find two-sided confidence intervals for σ_1^2/σ_2^2, using MINITAB we proceed as follows:

1. From the Menu bar, select **Stat** > **Basic Statistics** > **2-Variance**. This prompts a dialog box **Two-Sample Variance** to appear on the screen.
2. Select an appropriate option (Both samples are in one column. Each sample is in its own column, sample standard deviations or sample variances) from the pull down menu that appears on the Two-Sample Variance dialog box.
3. Check **Options**, which prompts another dialog box to appear. Select either variance or standard deviation ratio from the pull down menu next to **Ratio**. Enter the desired confidence level in the box next to **Confidence level**, set **Hypothesized ratio** to 1 and in the next to **Alternative hypothesis** select **Ratio ≠ hypothesized ratio** (**Ratio < hypothesized ratio** or **Ration > hypothesized ratio** for one sided confidence intervals). Click **OK** in each of the two dialog boxes. The MINITAB output shows up in the Session window as given below:



Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using F
0.866667	(0.503, 1.545)

Ratio of Variances

Estimated Ratio	95% CI for Ratio using F
0.751111	(0.253, 2.386)

USING R

We can use ‘qf(p, df1, df2, ncp, lower.tail = TRUE)’ function in R to get the F quantiles in the confidence interval calculations as follows.

```
#Assign variables
alpha = 0.05; n1 = 13; n2 = 16; S1 = 6.5; S2 = 7.5

#To obtain the two-sided confidence interval for σ_1^2/σ_2^2
CI = c((S1^2/S2^2)*(1/qf(1-alpha/2,n1-1, n2-1)),(S1^2/S2^2)*
      (1/qf(alpha/2,n1-1, n2-1)))
CI #R output
[1] 0.2534727, 2.3864311

#To obtain the two-sided confidence interval for σ_1/σ_2
sqrt(CI) #R output
[1] 0.5034607, 1.5448078
```

PRACTICE PROBLEMS FOR SECTIONS 8.5 AND 8.6

1. A sample of size 12 from a population assumed to be normal with unknown variance σ^2 yields $S^2 = 86.2$. Determine 95% confidence limits for σ^2.
2. Tensile strengths were measured for 15 test pieces of cotton yarn randomly taken from the production of a given spindle. The value of S for this sample of 15 tensile strengths is found to be 11.2 lb. Find 95% confidence limits for the standard deviation σ of the population. It is assumed that the population is approximately normal.
3. A sample of 220 items turned out during a given week by a certain process has an average weight of 2.57 lb and standard deviation of 0.57 lb. During the next week, a different lot of raw material was used, and the average weight of 220 items produced that week was 2.66 lb, and the standard deviation 0.48 lb. Assume normality.
   (a) Construct a 95% confidence interval for the ratio of the population variances. (Hint: You need to interpolate in the $F_{m,m,0.05}$ table or use one of the software packages.)
   (b) On the basis of the interval calculated in (a), determine a 95% confidence interval for the differences of the population mean weights.
4. Two types of tires are tested with the following results:

$$\text{Type A: } n_1 = 121,\ \bar{X}_1 = 27,\ 465 \text{ miles},\ S_1 = 2500 \text{ miles}$$

$$\text{Type B: } n_2 = 121,\ \bar{X}_2 = 29,\ 527 \text{ miles},\ S_2 = 3000 \text{ miles}$$

   (a) Assuming normality, find a 99% confidence interval for σ_1^2/σ_2^2.
   (b) On the basis of the interval found in (a), find a 99% confidence interval for $\mu_1 - \mu_2$.

5. A random sample from a normal population with unknown mean μ and variance σ^2 yields the following summary statistics: $n = 25$, $\bar{X} = 28.6$, $S = 5.0$
   (a) Find a 95% confidence interval for σ^2 and a 95% confidence interval for σ.
   (b) Find 95% one-sided lower and upper confidence limits for σ, and compare these limits with the limits you found in the two-sided 95% confidence interval for σ. Comment on why these limits are different.
6. A new catalyst is used in 10 batches of a chemical production process. The final yield of the chemical in each batch produce the following data:

27	28	25	30	27	26	25	27	30	30

   Assuming that chemical yields are normally distributed with mean μ_1 and variance σ_1^2, find a 95% confidence interval for the population variance. Find a one-sided 99% lower and upper one-sided confidence interval for the population variance σ_1^2.
7. In Problem 6, the chemical production from the last 15 batches in which the existing catalyst was used is:

31	33	28	28	30	31	30	32	30	33	35	34	28	32	33

   Assuming that these chemical yields are normally distributed with mean μ_2 and variance σ_2^2, find a 95% confidence interval for the population variance, say σ_2^2. Find a one-sided 99% lower and upper one-sided confidence interval for the population variance σ_2^2.
8. Using the information and the data provided in Problems 6 and 7, compute a 95% confidence interval for the ratio σ_1^2/σ_2^2 of the two variances.

8.7 POINT AND INTERVAL ESTIMATORS FOR THE PARAMETERS OF BINOMIAL POPULATIONS

8.7.1 One Binomial Population

Suppose that we are sampling from a Bernoulli population with probability mass function given by

$$f(x; p) = p^x (1-p)^{1-x}; \quad x = 0, 1 \tag{8.7.1}$$

We saw in Chapter 4 that the mean μ and variance σ^2 of this population are given by

$$\mu = p \text{ and } \sigma^2 = p(1-p) \tag{8.7.2}$$

respectively.

Now, if $X_1, \ldots, X_n$ is a random sample of n independent observations on X, whose probability function is given by Equation (8.7.1), then

$$T = X_1 + \cdots + X_n$$

has mean and variance $n\mu$ and $n\sigma^2$, respectively. That is,

$$E(T) = np \text{ and } Var(T) = np(1-p) \tag{8.7.3}$$

Now, if we denote the statistic T/n by $\hat{p}$, recall that $\hat{p}$ is the MLE of p, as seen in Example 8.2.5. From Equation (8.7.3), we have

$$E(\hat{p}) = p \text{ and } Var(\hat{p}) = p(1-p)/n \tag{8.7.4}$$

which is to say that $\hat{p}$ is an *unbiased point estimator* of p with variance $p(1-p)/n$.

Now recall from Theorem 5.7.1 that for large n ($np > 5$ and $n(1-p) > 5$) we can write, to good approximation,

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim Z \tag{8.7.5}$$

where Z is distributed as the standard normal variable $N(0,1)$, so that $(\hat{p}-p)/\sqrt{p(1-p)/n}$ is a pivotal quantity for p. Note that since the sample size n is large, we may estimate $Var(\hat{p})$ by $\hat{p}(1-\hat{p})/n$ in the pivotal quantity, so that (n large) we have

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{\alpha/2}\right) = 1-\alpha \tag{8.7.6}$$

Solving these inequalities, we may write that to good approximation (n large) that

$$P(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}) = 1-\alpha \tag{8.7.7}$$

Thus, from Equation (8.7.7), the $100(1-\alpha)\%$ confidence interval for p follows as:

$100(1-\alpha)\%$ *confidence interval for the population proportion* p:

$$(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}) \tag{8.7.8}$$

Lower and upper $100(1-\alpha)\%$ one-sided confidence intervals for p are given below by Equations (8.7.9a) and (8.7.9b), respectively, for large n:

One-sided $100(1-\alpha)\%$ *lower and upper confidence intervals for the population proportion* p:

$$(\hat{p} - z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n},\ 1) \tag{8.7.9a}$$

$$(0,\ \hat{p} + z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n}) \tag{8.7.9b}$$

Example 8.7.1 (Large sample confidence interval for binomial parameter p) *A random sample of 400 computer chips is taken from a large lot of chips and 50 of them are found to be defective. Find a 95% confidence interval for p the proportion of defective chips contained in the lot.*

Solution: From the information given, we have $n = 400,\ T = 50,\ \alpha = 0.05,\ \alpha/2 = 0.025,$ $z_{\alpha/2} = 1.96,\ \hat{p} = T/n = 50/400 = 0.125$

Substituting these values in Equation (8.7.8), we have

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = 0.125 - 1.96\sqrt{\frac{(0.125)(0.875)}{400}} = 0.0926$$

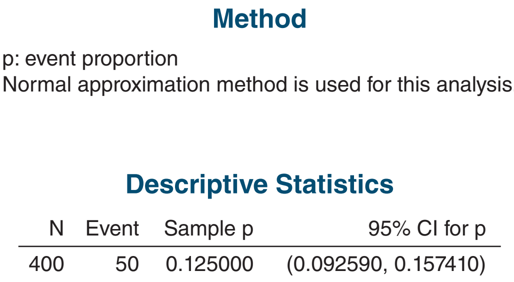
$$\hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = 0.125 + 1.96\sqrt{\frac{(0.125)(0.875)}{400}} = 0.1574$$

That is, a 95% confidence interval for p is (0.0926, 0.1574).

Interpretation: We are 95% confident that the population proportion of defective chips is between 9.26% and 15.74%.

MINITAB

1. Select **Stat** > **Basic Statistics** > **1 Proportion**; this prompts a dialog box titled **One-Sample Proportion** to appear on the screen.
2. Select **Summarized Data** from the pull down menu and enter the number of events (successes) and the number of trials in the appropriate boxes.
3. Check **Options**, which prompts another dialog box to appear. Enter 0.95, or the desired **Confidence level** in the box next to Confidence level and from the **Alternative hypothesis** option select **Proportion ≠ hypothesized proportion**, **Proportion < hypothesized proportion**, or **Proportion > hypothesized proportion** depending on whether a two-sided, one-sided upper, or one-sided lower confidence interval is sought, and select the method (Exact or Normal approximation) from the pull down menu next to **Method**.
4. Click **OK** in each of both the dialog boxes. The MINITAB output for Normal approximation shows up in the Session window as given below.



```
                    Method

p: event proportion
Normal approximation method is used for this analysis.


             Descriptive Statistics

  N  Event  Sample p          95% CI for p
400     50  0.125000  (0.092590, 0.157410)
```

USING R

The following manual R code can be used to obtain two-sided (normal approximation based) confidence interval.

```
#Assign variables
alpha = 0.05; T = 50; n = 400; phat = T/n


#To obtain the two-sided confidence interval for p
c(phat-qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n),phat-qnorm(alpha/2)*sqrt(phat*(1-phat)/n))
#R output
[1] 0.09259014, 0.15740986
```

8.7.2 Two Binomial Populations

Often we are interested in finding a confidence interval for the difference of the two population proportions. For example, we may be interested in estimating the true difference $(p_1 - p_2)$ of the failure rate of a product manufactured by two independent companies. One way to know which company's product is better is to find a confidence interval for $(p_1 - p_2)$ with a desired confidence coefficient.

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be random samples of sizes n_1 and n_2 from two independent Bernoulli populations with parameters p_1 and p_2, respectively. Then, we know that

$$\hat{p}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1},\ \hat{p}_2 = \frac{\sum_{j=1}^{n_2} X_{2j}}{n_2} \tag{8.7.10}$$

are unbiased estimators of p_1 and p_2, respectively. Therefore, $(\hat{p}_1 - \hat{p}_2)$ is an unbiased estimator of $(p_1 - p_2)$. Moreover, for large sample size $(n_1\hat{p}_1 > 5,\ n_1(1-\hat{p}_1) > 5)$ and $(n_2\hat{p}_2 > 5,\ n_2(1-\hat{p}_2) > 5)$, we know that $\hat{p}_1$ and $\hat{p}_2$ are approximately normally distributed with mean p_1 and p_2 and variance $p_1(1-p_1)/n_1$ and $p_2(1-p_2)/n_2$, respectively.

Now, from the result of Theorem 7.3.8, it follows that $(\hat{p}_1 - \hat{p}_2)$ is approximately normally distributed with mean $(p_1 - p_2)$ and variance $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$. That is,

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \tag{8.7.11}$$

is approximately distributed as the distribution of a standard normal $N(0,1)$. Thus, using the pivotal quantity in Equation (8.7.11), for large n we may write, to a good approximation,

$$P\left(-z_{\alpha/2} \le \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \le z_{\alpha/2}\right) = 1 - \alpha$$

or

$$P\left(-z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le (\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)\right.$$
$$\left.\le z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) = 1 - \alpha$$

Hence

$$P\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le (p_1 - p_2) \right.$$
$$\left. \le (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) = 1 - \alpha$$

Note that the quantity $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$ in both the lower and upper confidence limits is unknown, since p_1 and p_2 are not known. Also note that this quantity is the standard error of $(\hat{p}_1 - \hat{p}_2)$. Thus, we estimate the standard error of $(\hat{p}_1 - \hat{p}_2)$ by entering the expression for the standard error of $(\hat{p}_1 - \hat{p}_2)$, replacing p_1 and p_2 by $\hat{p}_1$ and $\hat{p}_2$, respectively. Now, a confidence interval for $(p_1 - p_2)$ can be derived for the large samples case (with confidence coefficient $(1-\alpha)$) and is given below by Equation (8.7.12), assuming n_1 and n_2 large.

$100(1-\alpha)\%$ *confidence interval for the difference of two population proportions* $(p_1 - p_2)$.

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ (\hat{p}_1 - \hat{p}_2) \right.$$
$$\left. + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) \tag{8.7.12}$$

Note: The lower limit cannot be less than -1 and the upper limit cannot be greater than 1.

Lower and upper one-sided confidence intervals for $(p_1 - p_2)$ with confidence coefficient $(1-\alpha)$ are given below by Equations (8.7.13a) and (8.7.13b), respectively.

One-sided $100(1-\alpha)\%$ *upper and lower confidence intervals for the difference of two population proportions* $(p_1 - p_2)$ are, respectively

$$\left(-1,\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) \tag{8.7.13a}$$

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ 1\right) \tag{8.7.13b}$$

Example 8.7.2 (A large sample confidence interval for the difference of two binomial parameters $(p_1 - p_2)$) *Two companies A and B claim that a new type of light bulb has a lifetime of more than 5000 hours. In a random sample of 400 bulbs manufactured by* *company A, 60 bulbs burned out before the claimed lifetime, and in another random sample of 500 bulbs manufactured by company B, 100 bulbs burned out before the claimed lifetime. Find a point estimate and a 95% confidence interval for the true value of the difference $(p_1 - p_2)$, where p_1 and p_2 are the proportion of the bulbs manufactured by company A and company B, respectively, that burn out before the claimed lifetime of 5000 hours.*

Solution: From the information given, we have

$$\hat{p}_1 = 60/400 = 3/20 \text{ and } \hat{p}_2 = 100/500 = 1/5$$

Thus, the point estimate of $(p_1 - p_2)$ is

$$\hat{p}_1 - \hat{p}_2 = 3/20 - 1/5 = -1/20 = -0.05$$

Since the sample sizes are large, to find a 95% confidence interval for $(p_1 - p_2)$, we use interval (8.7.12). Now substituting the values of $\hat{p}_1$ and $\hat{p}_2$, and using $z_{0.025} = 1.96$ since $\alpha = 0.05, \alpha/2 = 0.025$, we obtain

$$\begin{aligned} LCL &= \left(\frac{3}{20} - \frac{1}{5}\right) - 1.96\sqrt{\frac{\frac{3}{20}(1-\frac{3}{20})}{400} + \frac{\frac{1}{5}(1-\frac{1}{5})}{500}} \\ &= -0.05 - 0.04953 = -0.09953 \\ UCL &= \left(\frac{3}{20} - \frac{1}{5}\right) + 1.96\sqrt{\frac{\frac{3}{20}(1-\frac{3}{20})}{400} + \frac{\frac{1}{5}(1-\frac{1}{5})}{500}} \\ &= -0.05 + 0.04953 = -0.00047 \end{aligned}$$

Thus, a two-sided 95% confidence interval for $(p_1 - p_2)$ is (–0.09953,–0.00047).

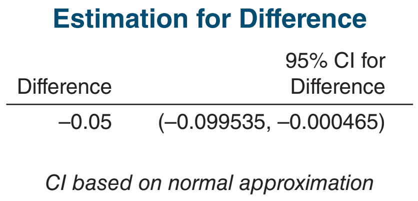
Since the confidence interval for $(p_1 - p_2)$ does not contain 0, we can support the statement that the quality of bulbs manufactured by the companies is different, at confidence level 95%.

Further, as both the confidence limits are negative, we can conclude at the 5% level of significance that the quality of bulbs manufactured by company B is inferior to those manufactured by company A.

MINITAB

1. Select **Stat** > **Basic Statistics** > **2 Proportions**. This prompts a dialog box titled **Two-Sample Proportion** to appear on the screen.
2. Select **Summarized Data** from the pull down menu and enter the number of events (successes) and the number of trials in the appropriate boxes. (Here "success" means that a bulb failed before 5000 h.)
3. Check **Options**, which prompts another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level**, enter 0 in the box next to **Hypothesized difference** and from the **Alternative hypothesis** option select **Difference ≠ hypothesized difference**, **Difference < hypothesized**

difference, or **Difference > hypothesized difference** depending on whether a two-sided, one-sided upper, or one-sided lower confidence interval is sought and select **Use the pooled estimate of the proportion** from the pull down menu next to **Test method**. Click **OK** in each of both the dialog boxes. The MINITAB output shows up in the Session window as given below.



```
            Estimation for Difference

                            95% CI for
Difference                  Difference
--------------------------------------------
     –0.05      (–0.099535, –0.000465)

         CI based on normal approximation
```

USING R

The R function ‘prop.test()’ can be used to obtain the required two-sided pooled normal approximation based confidence interval as shown in the following R code.

```
Test = prop.test(x = c(60, 100), n = c(400, 500), conf.level = 0.95, correct = FALSE)
#‘x’ and ‘n’ indicate the number of successes (burned out) and the total
number of trials in each sample, respectively, and the option ‘correct = FALSE’
indicates that the continuity correction should not be applied.


Test$conf.int #R output
[1] -0.0995351574, -0.0004648426
attr(,“conf.level”)
[1] 0.95
```

PRACTICE PROBLEMS FOR SECTION 8.7

1. A random sample of 500 individuals contains 200 that wear eyeglasses. Find a 95% confidence interval for p, the proportion of people in the population wearing glasses.
2. A new coin is tossed 50 times, and 20 of the tosses show heads and 30 show tails. Find a 99% confidence interval for the probability of obtaining a head when this coin is tossed again.
3. A sample of 100 consumers showed 16 favoring Brand X of orange juice. An advertising campaign was then conducted. A sample of 200 consumers surveyed after the campaign showed 50 favoring Brand X. Find a 95% confidence interval for the difference in proportions of the population favoring Brand X before and after the advertising campaign. Comment on your result.
4. An orange juice manufacturing company uses cardboard cans to pack frozen orange juice. A six sigma black belt quality engineer found that some of the cans supplied by supplier A do not meet the specifications, as they start leaking at some point. In a random sample of 400 packed cans, 22 were leaking. Find a 95% confidence interval

for the fraction of leaking cans. Find one-sided 95% lower and upper confidence intervals for the fraction of leaking cans.

5. During the past 50 years, more than 10 serious nuclear accidents have occurred worldwide. Consequently, the US population is divided over building more nuclear plants. Suppose two US states decide to determine the percentage of their population that would like to have a nuclear plant in their state. In a random sample of 800 persons from one state, only 40 favored having a nuclear plant; a random sample of 600 persons from another state showed 50 persons favored having a nuclear plant. Find a 95% confidence interval for the difference between the percentages of persons who favor a nuclear plant in their state.
6. A random sample of 800 homes from a metropolitan area showed that 300 of them are heated by natural gas. Determine a 95% confidence interval for the proportion of homes heated by natural gas.
7. In a manufacturing plant, two machines are used to produce the same mechanical part. A random sample of 225 parts produced by machine 1 showed that 12 of the parts are nonconforming, whereas a random sample 400 parts produced by the machine 2 showed that 16 of the parts are nonconforming. Construct a 95% confidence interval for the difference of the proportions of nonconforming parts produced by the two machines.
8. An instructor is interested in determining the percentage of the students who prefer multiple choice questions versus open-ended questions in an exam. In a random sample of 100 students, 40 favored multiple choice questions. Find a 95% confidence interval for the proportion of students who favor multiple-choice questions in an exam.
9. In a random sample of 500 voters interviewed across the nation, 200 criticized the use of personal attacks in an election campaign. Determine a 90% confidence interval for the proportion of voters who criticize the use of personal attacks in the election campaign.

8.8 DETERMINATION OF SAMPLE SIZE

In this section, we discuss the determination of the sample size needed to estimate a parameter θ when the margin of error E is known. Here, θ may be the population mean μ, or the difference of the two population means $\mu_1 - \mu_2$, or the population proportion p, or the difference of the two population proportions $p_1 - p_2$, and so on.

For example, let $X_1, \ldots, X_n$ be a random sample from a population with probability distribution $f(x, \theta)$ where θ is an unknown parameter. Let $\hat{\theta} = \varphi(X_1, \ldots, X_n)$ be an estimate of θ, where, clearly, one cannot expect $\hat{\theta}$ to be exactly equal to the true value of θ. The difference between $\hat{\theta}$ and θ is the *error of estimation*. The maximum value of the error of estimation is called the *margin of error* or *bound on error of estimation*, where the margin of error of estimation, denoted by E (see Equation (8.2.6)) for a confidence level probability $(1 - \alpha)$, is given by

$$E = z_{\alpha/2}\sigma_{\hat{\theta}}$$

Note that E is equal to half the width of the confidence interval for θ with confidence coefficient $(1 - \alpha)$. Suppose that the size of the margin of error E is predetermined; we then would like to find the sample size needed to attain this value of the margin of error.

8.8.1 One Population Mean

Let $\theta = \mu$. Then, the margin of error with probability $(1-\alpha)$ is given by

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where here σ is the known population standard deviation. Squaring both sides of the equation and doing some algebraic manipulations, we obtain

Sample size for estimating a population mean μ with margin of error E and with probability $(1-\alpha)$:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{E^2} \tag{8.8.1}$$

Example 8.8.1 (Determination of sample size n) *A manufacturing engineer wants to estimate the average number of defective parts produced by a machine during each shift. An earlier study on a similar machine shows that the number of defective parts produced by the machine varies from shift to shift with a standard deviation equal to 12. How large a sample should the engineer take so that, with 95% probability, the estimate of the average number of defective parts produced by the machine is within three parts of the true value of μ, the average number of defective parts produced by the machine in each shift?*

Solution: From the information given, we have $z_{\alpha/2} = z_{0.025} = 1.96,\ \sigma = 12,\ E = 3$
Thus, the desired sample size is

$$n = \frac{z_{\alpha/2}^2\sigma^2}{E^2} = \frac{(1.96)^2(12)^2}{3^2} = 61.46$$

Therefore, the engineer should take a sample of size 62 to achieve his/her goal. In order to be certain the value of E attained is as described, the value of n should always be rounded up.

8.8.2 Difference of Two Population Means

When $\theta = \mu_1 - \mu_2$ is of interest, consider the case $n_1 = n_2 = n$, n to be determined. The margin of error with probability $(1-\alpha)$ is given by

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}} = z_{\alpha/2}\frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{n}}$$

Now, by squaring both sides of this equation and doing some algebraic manipulations, we have the sample size for estimating the difference of two population means $\mu_1 - \mu_2$ with margin of error E:

Sample size n ($n = n_1 = n_2$) *for estimating the difference of two population means* $\mu_1 - \mu_2$ *with margin of error E and with probability* $(1 - \alpha)$:

$$n = \frac{z^2_{\alpha/2}(\sigma_1^2 + \sigma_2^2)}{E^2} \tag{8.8.2}$$

Here, σ_1^2 and σ_2^2 are the known variances of populations under consideration.

Example 8.8.2 (Determination of a common sample size n, $n_1 = n_2 = n$) *Suppose that an experimenter wishes to estimate the difference between two population means* μ_1 *and* μ_2*, and suppose that we know* $\sigma_1 = 2.0$ *and* $\sigma_2 = 2.5$*. How large a sample of common size should be taken from each population so that, with probability 99%, the estimate for* $\mu_1 - \mu_2$ *is within 1.2 units of the true value of* $\mu_1 - \mu_2$*?*

Solution: From the information given, we have $1 - \alpha = 0.99, \alpha/2 = 0.005$, and

$$z_{\alpha/2} = z_{0.005} = 2.575, \sigma_1 = 2.0, \sigma_2 = 2.5, E = 1.2$$

Thus, from Equation (8.8.2), the desired common sample size is

$$n = \frac{(2.575)^2(2^2 + (2.5)^2)}{(1.2)^2} = 47.197 \approx 48$$

In practice, it is quite common for the population variance (or variances) not to be known. In such cases, we replace the population variance by the sample variance. It is interesting here that we do not know the population variance, but we have to find the sample variance for which we need to have a sample. To have a sample, we must know the sample size that we want to find. This can become a vicious circle. To solve this problem, we take one of two possible approaches:

1. We use some existing data (or data collected from a pilot study) to calculate an estimate of the sample variance. Then, we use that estimate of the sample variance in our determination of the sample size n, replacing in Equation (8.8.2) with $\hat{\sigma}$.
2. We take a preliminary small sample, say of size n_1, to calculate the value of the sample variance. Then, we use this value of the sample variance to determine the sample size n. Since we already have a sample size n_1, we take another supplemental sample of size $n - n_1$ and then combine the two samples in order to get a full sample of size n.

8.8.3 One Population Proportion

Let p be the population proportion. In this case, the margin of error E with probability $(1 - \alpha)$ is given by

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Now squaring both sides of the equation and doing some algebraic manipulation, we obtain

> *Sample size for estimating population proportion p with margin of error E and with probability* $(1-\alpha)$:
>
> $$n = \frac{z_{\alpha/2}^2 p(1-p)}{E^2} \tag{8.8.3}$$

where, p is unknown, and we handle this aspect as illustrated in the following example.

Example 8.8.3 (Determination of sample size n) *Suppose a random sample of eligible voters from some district is selected to estimate the proportion p of voters who favor the incumbent candidate. How large a sample should be taken in order to estimate p with a margin of error of 3%, and with 95% probability?*

Solution: From the information available to us, we have

$$z_{\alpha/2} = z_{0.025} = 1.96,\ E = 3\% = 0.03.$$

Since we have no prior information about p, in order to make certain that our margin of error is no more than 3%, we choose to use $p = 0.5$. This choice maximizes Equation (8.8.3); that is, it gives us the largest possible sample needed to attain the given margin of error. This is because $p(1-p)$ attains its maximum value, 1/4, at $p = 0.5$. Thus, using Equation (8.8.3), the sample size is

$$n = \frac{(1.96)^2(0.5)(1-0.5)}{(0.03)^2} = 1067.11 \approx 1068$$

so we would use a sample of size 1068.

8.8.4 Difference of Two Population Proportions

Let the difference of two Bernoulli population proportions be $p_1 - p_2$. Suppose that the sample sizes taken from the two populations are equal, that is, $n_1 = n_2 = n$. Then the margin of error E where estimating $p_1 - p_2$ is given by

$$E = z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$

Again, taking the square on both sides and doing some algebraic manipulations, we find that the desired common sample size n needed for a margin of error no greater than E with probability $(1-\alpha)$ is as follows:

> *Sample size for estimating the difference of two population proportions* $p_1 - p_2$ *with margin of error E for n* $(n = n_1 = n_2)$ *and with probability* $(1-\alpha)$:
>
> $$n = \frac{z_{\alpha/2}^2[p_1(1-p_1) + p_2(1-p_2)]}{E^2} \tag{8.8.4}$$

Example 8.8.4 (Determination of a common sample size n, $n_1 = n_2 = n$) *A marketing specialist in a car manufacturing company wants to estimate the difference between the proportion of those customers who prefer a domestic car and those who prefer an imported car. How large a sample should the specialist take from those who prefer domestic cars and those who prefer imported cars in order to have a margin of error of 2.5% with probability 99%? It is known that in previous visit to the car show room, 60% of the customers preferred domestic and 40% preferred imported cars.*

Solution: From the information available to us, we have $\hat{p}_1 = 0.6$, $\hat{p}_2 = 0.4$, $z_{\alpha/2} = z_{0.005} = 2.575$, $E = 2.5\% = 0.025$

Substituting these values appropriately in Equation (8.8.4), we obtain

$$n = \frac{(2.575)^2[0.6 \times 0.4 + 0.4 \times 0.6]}{(0.025)^2} = 5092.32 \approx 5093$$

so we would use sample sizes 5093 from each of the two groups.

PRACTICE PROBLEMS FOR SECTION 8.8

1. A manufacturer of electric bulbs wants to find a 98% confidence interval for the lifetime of its new family of bulbs. How large a sample should be selected if the company seeks a confidence interval for μ no more than 20 hours in width? Assume that previous experience has shown that the standard deviation of the lifetime of these bulbs is 30 hours.
2. In Problem 1, what should the sample size be if the confidence coefficient is decreased from 98% to 95%? What should the sample size be if the standard deviation of the lifetime of these bulbs is actually 50 (instead of 30) hours?
3. Assuming that the sample sizes selected from two populations for estimating the difference of two population proportions are equal, find the largest possible common sample size that would be needed in order to be 95% confident that the margin of error is (a) 0.04, (b) 0.035, (c) 0.03.
4. Determine how large a sample size is required so that we are 90% confident that the margin of error in estimating a population mean is (a) 20, (b) 35, (c) 45, (d) 65. Assume that it is known from many earlier studies that the population standard deviation is 40.
5. Two independent random samples of size n each are selected from two binomial populations. If we wish to estimate the difference between the two population proportions p_1 and p_2 correct to within 0.04 with probability equal to 95%, how large should n be? Assume that (a) we have no prior information on p_1 and p_2 and (b) some historical data indicate that $p_1 \approx 0.35$ and $p_2 \approx 0.47$.
6. Two independent random samples, each of size n, are selected from two populations. If we wish to estimate the difference between the two population means correct to within 2.5 with probability equal to 95%, how large should n be? Assume that using some historical data, we found that $S_1^2 = 37$ and $S_2^2 = 29$.
7. A study is being proposed to estimate the proportion of residents in a certain town who favor the construction of a new high school. How large a sample is needed to have 98% confidence that the estimate is within 0.03 of the true proportion of residents favoring the construction of a new high school?

8. How large a sample is needed in Problem 10 of Section 8.4, if one wants to be 95% confident that the estimate for the difference between the two means is within 0.5 of the true difference?
9. A gynecologist wishes to estimate the mean age at which women are first diagnosed with cervical cancer. To this end, she wishes to construct a 98% confidence interval for this age, which is six years wide. If the population standard deviation is known to be eight years, how large a sample should be taken?
10. A pediatric researcher is interested in estimating the difference between the head circumferences of newborn babies in two populations. How large should the samples be taken if she wants to construct a 95% confidence interval for the difference between the head circumferences that is 2 cm wide? Assume that the two population standard deviations are known to be 1.5 and 2.5 cm and that equal-sized samples are to be taken.

8.9 SOME SUPPLEMENTAL INFORMATION

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

8.10 A CASE STUDY

Case Study[1] During the qualification of a new microchip product, the LMV9234, at National Semiconductor, three nominal prototype lots of 10 wafers each are run through the manufacturing line using the standard process of record. Each wafer has nine electrical sites at which over 70 critical electrical parameters for transistors, resistors, diode, and capacitors are tested at the end of the line. The threshold voltage for the NMOS transistors, V_T, is a critical parameter of interest that is tested at the end of the line.

The product engineer for this new product would like to test if the mean threshold voltage for the NMOS transistors of the first lot for the LMV9234 is different from that of the historical threshold voltage for the last 180 days running in the fabrication facility. For this lot, five electrical sites from each of the 10 wafers were tested (A wafer is a thin slice of semiconductor material, such as silicon crystal, used in manufacturing of integrated circuit and other micro-devices; see Figure 8.10.1.) The V_T data are summarized in Table 8.10.1. Compute the appropriate test statistic and examine whether the V_T is different from the historical value for the threshold voltage, V_T, of 0.609055 volts for the last 180 days. Find a 90%, 95%, and 99% confidence interval for V_T. Analyze the results of the case study. Prepare a short report summarizing your conclusions. The data for this case study (Case Study 8.10.1) are available on the book website: www.wiley.com/college/gupta/statistics2e.

8.11 USING JMP

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

[1] Source: Major Integrated Circuit Manufacturer.

Figure 8.10.1 An etched silicon wafer.

Review Practice Problems

1. In estimating the mean of a normal distribution $N(\mu, \sigma^2)$ having known standard deviation σ by using a confidence interval based on a sample of size n, what is the minimum value of n in order for the 99% confidence interval for μ to be of length not greater than L?

2. Two machines A and B are packaging 8-oz boxes of corn flakes. From past experience with the machines, it is assumed that the standard deviations of weights of the filling from the machines A and B are 0.04 oz and 0.05 oz, respectively. One hundred boxes filled by each machine are selected at random and the following is found:

$$\text{Machine } A\text{: } n_A = 100,\ \bar{X}_A = 8.18 \text{ oz}$$
$$\text{Machine } B\text{: } n_B = 100,\ \bar{X}_B = 8.15 \text{ oz}$$

   Find a 99% confidence interval for $\mu_A - \mu_B$, the difference of the means of populations of weights of fillings produced by machines A and B.

3. Four determinations of the pH of a certain solution are 7.90, 7.94, 7.91, and 7.93. Assuming normality of determinations with mean μ, standard deviation σ, find (a) 99% confidence limits for μ and (b) 95% confidence limits for σ.

4. Ten determinations of percentage of water in a methanol solution yield $\bar{X} = 0.552$ and $S = 0.037$. If μ is the "true" percentage of water in the methanol solution, assuming

normality, find (a) a 90% confidence interval for μ and (b) a 95% confidence interval for σ.

5. A clock manufacturer wants to estimate the variability of the precision of a certain type of clock being manufactured. To do this, a sample of eight clocks is run for exactly 48 hours. The number of seconds each clock is ahead or behind after the 48 hours, as measured by a master clock, is recorded, the results being $+6, -4, -7, +5, +9, -6, -3, +2$. Assuming the time recorded by clocks of this kind to be normally distributed, find:
   (a) A 95% confidence interval for the mean time recorded by this type of clock after 48 hours.
   (b) A 95% confidence interval for σ, the standard deviation of the time recorded by this type of clock after 48 hours.

6. The following data give the yield point (in units of 1000 psi) for a sample of 20 steel castings from a large lot:

64.5	66.5	67.5	67.5	66.5	65.0	73.0	63.5	68.5	70.0
71.0	68.5	68.0	64.5	69.5	67.0	69.5	62.0	72.0	70.0

Assuming that the yield points of the population of castings to be (approximately) normally distributed, find:
   (a) A 95% confidence interval for μ, the mean yield point of the lot
   (b) A 95% confidence interval for σ, the standard deviation of the yield points of the lot.

7. In firing a random sample of nine rounds from a given lot of ammunition, the tester finds that the standard deviation of muzzle velocities of the nine rounds is 38 ft/s. Assuming that muzzle velocities of rounds in this lot are (approximately) normally distributed, find 95% confidence limits for the standard deviation σ of the muzzle velocity of the lot.

8. A sample of four tires is taken from a lot of brand-A tires. Another sample of four tires is taken from a lot of brand-B tires. These tires are tested for amount of wear, for 24,000 miles of driving on an eight-wheel truck, when the tires are rotated every 1000 miles. The tires are weighed before and after the test. The loss in weight, expressed as the percentage of initial weight, is used as a measure of wear. For the sample of brand-A tires, it is found that $\bar{X}_A = 18.0, S_A = 1.3$; for brand-B tires, $\bar{X}_B = 19.4, S_B = 1.5$. Assuming (approximately) normal distributions having equal variances for percent wear of tires of each brand under these test conditions, find a 95% confidence interval for $\mu_A - \mu_B$.

9. The following data give the Vickers hardness number of ten shell castings from company A and of ten castings from company B:

Company A (Hardness):	66.3	64.5	65.0	62.2	61.3	66.5	62.7	67.5	62.7	62.9
Company B (Hardness):	62.2	67.5	60.4	61.5	64.8	60.9	60.2	67.8	65.8	63.8

Find a 90% confidence interval for $\mu_A - \mu_B$, assuming (approximate) normality and that $\sigma_A^2 = \sigma_B^2 = \sigma^2$.

10. Resistance measurements (in ohms) are made on a sample of four test pieces of wire from lot A and five from lot B, with the following results given:

Lot A Resistance (ohms):	0.143	0.142	0.143	0.137	
Lot B Resistance (ohms):	0.140	0.142	0.136	0.138	0.14

If μ_1 and μ_2 are the mean resistances of the wire in lots A and B and assuming that possible measurements from lot A and possible measurements from lot B have normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where μ_1, μ_2 and σ^2 are unknown, find a 95% confidence interval for $\mu_1 - \mu_2$. Use this interval to examine the statement that $\mu_1 = \mu_2$.

11. Breaking strengths (in psi) are observed on a sample of five test pieces of type-A yarn and nine test pieces of type-B yarn, with the following results:

Type A (psi):	93	94	75	84	91				
Type B (psi):	99	93	99	97	90	96	93	88	89

Assuming normality of breaking strengths for each type of yarn and that the population variances of breaking strengths are equal in the two populations, find 99% confidence limits for $\mu_A - \mu_B$, where μ_A and μ_B are the population means.

12. Determinations of atomic weights of carbon from two preparations I and II yield the following results:

Preparation I:	12.0072	12.0064	12.0054	12.0016	12.0077
Preparation II:	11.9583	12.0017	11.9949	12.0061	

Assuming (approximate) normality, find a 95% confidence interval for $\mu_1 - \mu_2$, the difference of the true means of determinations made from preparations I and II. (*Note*: We have not assumed equality of variances.)

13. An experiment to determine the viscosity of two different types of gasoline (leaded versus nonleaded) give the following results:

$$\text{Leaded: } n_1 = 25,\ \bar{X}_1 = 35.84,\ S_1^2 = 130.4576$$

$$\text{Nonleaded: } n_2 = 25,\ \bar{X}_2 = 30.60,\ S_2^2 = 53.0604$$

Assuming normality, find a 95% confidence interval for the difference in viscosities. (*Note*: We have not assumed equality of variances.)

14. Two analysts A and B each make 10 determinations of percentage chlorine in a batch of polymer. The sample variances S_A^2 and S_B^2 turn out to be 0.5419 and 0.6065, respectively. If σ_A^2 and σ_B^2 are the variances of the populations of A's measurements and of B's measurements respectively, find 95% confidence limits for σ_A^2/σ_B^2.

15. For the data of Problem 8, find a 95% confidence interval for σ_A^2/σ_B^2, the ratio of the population variances. Comment on the assumption made in Problem 8, that the true value of the ratio is one.

16. In Problems 9, 10, and 11, the assumption was made that the ratio of the variances of the populations being sampled is 1. In each of these problems, find 95% confidence intervals for the ratio of the variances of the populations being sampled and comment on the assumption.

17. A large university wishes to estimate the mean expense account of the members of its staff who attend professional meetings. A random sample of 100 expense accounts yields a sample average of \$515.87 with a sample standard deviation of \$14.34. Find a 95% confidence interval for the mean expense account amount, say μ.

18. A new spinning machine is installed and 64 test pieces of yarn it produces are tested for breaking strength. The observed data yield a sample average of 4.8 and standard deviation of 1.2. Find a 99% confidence interval for the true value of the breaking strength.

19. The heights of 105 male students of university X chosen randomly yield an average of 68.05 inches and a sample standard deviation of 2.79 in. Find a 95% confidence interval for the mean of the heights of male students attending the (rather large) university X.

20. A sample of 70 employees selected at random from the employees of a large brewery yields an average disabled time of 41.8 hours during a fiscal year, with a standard deviation of 6.4 hours. Construct a 99% confidence interval for the mean disabled time of employees at this firm.

21. Two groups of judges are asked to rate the tastiness of a certain product. The results are as follows:

$$\text{Group 1: } n_1 = 121,\ \bar{X}_1 = 3.6,\ S_1^2 = 1.96$$
$$\text{Group 2: } n_2 = 121,\ \bar{X}_2 = 3.2,\ S_2^2 = 3.24$$

    (a) Find a 95% confidence interval for σ_1^2/σ_2^2, assuming normality.
    (b) On the basis of the interval found in (a), find a 95% confidence interval for $\mu_1 - \mu_2$.

22. A sample of 500 rounds of ammunition supplied by manufacturer A yields an average muzzle velocity of 2477 ft/s, with a standard deviation of 80 ft/s. A sample of 500 rounds made by another manufacturer, manufacturer B, yields an average muzzle velocity of 2422 ft/s, with a standard deviation of 120 ft/s. Find a 99% confidence interval for the differences in muzzle velocity of the bullets supplied by A and B. (First verify whether or not you can assume $\sigma_A^2 = \sigma_B^2$ on the basis of the previously mentioned sample evidence.)

23. A sample of 100 workers in one large plant took an average time of 23 minutes to complete a task, with a standard deviation of 4 minutes. In another large, but similar, plant, a sample of 100 workers took an average time of 25 minutes to complete the same task, with a standard deviation of 6 minutes.
    (a) Construct a 99% confidence interval for the ratio of the population variances, assuming normality.
    (b) On the basis of (a), determine a 99% confidence interval for the difference between the two population means.

24. A group of 121 students of a large university are retested on entrance to give an average score of 114 with a standard deviation of 19.6. Another group of 91 students

who have spent one year at this university are given the same test; they performed with an average score of 121 and standard deviation of 16.8. Assume normality.

(a) Find a 99% confidence interval for the ratio of the class variances.
(b) Using the results of (a), determine an appropriate 99% confidence interval for the difference of mean scores of the classes.

25. The effectiveness of two drugs is tested on two groups of randomly selected patients with the following results (in coded units):

$$\text{Group 1: } n_1 = 75,\ \bar{X}_1 = 13.540,\ S_1 = 0.476$$

$$\text{Group 2: } n_2 = 45,\ \bar{X}_2 = 11.691,\ S_2 = 0.519$$

Assume normality.

(a) Find a 95% confidence interval for $\sigma_1^2 = \sigma_2^2$.
(b) On the basis of the interval found in (a), find a 95% confidence interval for $\mu_1 - \mu_2$, the difference of the mean effectiveness of the two drugs.

26. The athletic department of a large school randomly selects two groups of 50 students each. The first group is chosen from students who voluntarily engage in athletics, the second group is chosen from students who do not engage in athletics. Their body weights are measured with the following results:

$$\text{Group 1: } n_1 = 50,\ \bar{X}_1 = 158.26 \text{ lb},\ S_1 = 7.08 \text{ lb}$$

$$\text{Group 2: } n_2 = 50,\ \bar{X}_2 = 151.47 \text{ lb},\ S_2 = 7.92 \text{ lb}$$

Assume normality.

(a) Find a 99% confidence interval for $\sigma_1^2 = \sigma_2^2$.
(b) Using (a), find a 99% confidence interval for $\mu_1 - \mu_2$. Comment.

27. Nine out of 15 students polled favored the holding of a demonstration on campus against “the war.” Using technology, find 95% confidence limits for the proportion of all students favoring this proposal.

28. A random sample of 60 voters selected at random from a large city indicate that 70% will vote for candidate A in the upcoming mayoral election. Find the 99% confidence interval for the proportion of voters supporting candidate A.

29. If $(X_1, \ldots, X_n)$ is a random sample of size n from $N(\mu, \sigma_0^2)$, where σ_0^2 is the known value of the population variance, find the maximum likelihood estimator of μ. Is the maximum likelihood estimator unbiased for μ? What is the distribution of the maximum likelihood estimator?

30. If $(X_1, \ldots, X_n)$ is a random sample of size n from $N(\mu_0, \sigma^2)$, where μ_0 is the known value of the population mean, find the maximum likelihood estimator of σ^2. Is the maximum likelihood estimator unbiased or not for σ^2? What is its distribution? What is its variance?

31. A manufacturing engineer wants to use the mean of a random sample of size 36 to estimate the average length of the rods being manufactured. If it is known that $\sigma = 1.5$ cm, find the margin of error at the 95% confidence level.

32. A past study indicates that the standard deviation of hourly wages of workers in an auto industry is \$4.00. A random sample of hourly wages of 49 workers yields an average of \$55.00. Find (a) a point estimate of population mean wage, (b) the standard error of the point estimator calculated in part (a), (c) the margin of error of the estimate at the 95% confidence level.

33. In a study of diameters of ball bearings manufactured by a newly installed machine, a random sample of 64 ball bearings is taken, yielding a sample mean of 12 mm and a sample standard deviation of 0.6 mm. Compute a 99% confidence interval for the population mean μ.

34. Two types of copper wires used in manufacturing electrical cables are being tested for their tensile strength. From previous studies, it is known that the tensile strengths of these wires are distributed with unknown means μ_1 and μ_2 but known standard deviations $\sigma_1 = 6.0$ psi and $\sigma_2 = 8.5$ psi, respectively. Two random samples, one of size $n_1 = 36$ of type I wire and another sample of size $n_2 = 49$ of type II wire yield sample means of 203 psi and 240 psi, respectively.
    (a) Determine a 90% confidence interval for the difference $\mu_1 - \mu_2$ of the two population means.
    (b) Determine a 99% confidence interval for the difference $\mu_1 - \mu_2$ of the two population means.

35. A recent study shows that the average annual incomes of cardiologists and gastroenterologists based on random samples of 100 and 121 are \$295,000 and \$305,000, respectively. Furthermore, these samples yield sample standard deviations of \$10,600 and \$12,800, respectively. Determine a 98% confidence interval for the difference $\mu_1 - \mu_2$ of the two population means.

36. Two samples of sizes $n_1 = 16$ and $n_2 = 25$ pieces of wool yarn are randomly taken from the production of two spindles and tested for tensile strength. These tests produce the following data:

Sample I:	12.28	8.54	11.31	9.06	10.75	10.96	12.12	8.14	10.75
	9.55	9.56	11.00	9.10	9.91	10.08	9.54		
Sample II:	12.89	11.35	13.15	13.84	10.86	13.45	13.19	11.21	12.07
	13.90	11.93	11.87	12.68	12.23	11.69	12.54	11.55	11.19
	12.36	12.82	13.12	13.07	11.86	11.65	11.96		

Assuming that the two populations are normally distributed with equal variances, find a 95% confidence interval for the difference $\mu_1 - \mu_2$ of the two population means.

37. At a certain university, the Electrical Engineering faculty decides to teach a course on robot intelligence using two different methods, one an existing method and the other a method using new strategies. The faculty randomly selects two sections of students who are scheduled to take this course. One section is taught using the existing method, while the other uses the new method. At the end of the semester, both sections are given the same test. Their test scores produce the following summary statistics:

$$n_1 = 49,\ \bar{X}_1 = 79,\ S_1^2 = 30$$

$$n_2 = 45,\ \bar{X}_2 = 86,\ S_2^2 = 40$$

(a) Find a 95% confidence interval for the difference $\mu_1 - \mu_2$ of the two population means.

(b) Find a 98% lower confidence interval for the difference $\mu_1 - \mu_2$ of the two population means and then find a 98% upper one-sided confidence interval for the difference $\mu_1 - \mu_2$ of the two population means.

38. A semiconductor company wants to estimate the fraction p of defective computer chips it produces. Suppose that a random sample of 900 chips has 18 defective chips. Find a point estimate of p. Next, find a 95% confidence interval for the proportion p of defective chips, and also find a one-sided 95% lower confidence interval for the proportion p of defective chips.

39. A company with two car-repair centers in Orange County, California, is interested in estimating the percentage of car owners who are very happy with the service they received at each center. In a random sample of 120 car owners who have their cars serviced at service center I, 72 are quite happy, while in another sample of 150 car owners who have their cars serviced at service center II, 110 are quite happy. Find a 95% confidence interval for the difference between the percentages of persons who are happy with the car service they have received at centers I and II.

40. Referring to Problem 38, how large a sample should be taken in order to be confident with probability 95% that the margin of error to estimate the fraction p of defective computer chips is 0.025?

41. Referring to Problem 40, find by how much the sample size increases or decreases if we are willing to increase the margin of error from 0.025 to 0.05 with the same probability of 0.95.

42. Suppose in Problem 37 that it had been agreed to take equal sample sizes from the two populations. What size samples should be selected so that we can be 99% confident that the margin of error in estimating the difference of two population means is no more than two points? Use the sample variances given in Problem 37 as estimates of σ_1^2 and σ_2^2.

43. Two textile mills are manufacturing a type of utility rope. A consumer group wants to test the tensile strength of the rope manufactured by these mills. A random sample of 10 pieces of rope manufactured by mill I results in a sample mean of $\bar{X}_1 = 850$ psi and a sample standard deviation of $S_1 = 30$ psi. A random sample of 15 pieces of rope manufactured by mill II results in a sample mean of $\bar{X}_2 = 880$ psi and a sample standard deviation of $S_2 = 38$ psi. Find a one-sided 95% upper confidence limit for the ratio σ_1^2/σ_2^2 of the two variances, assuming that the tensile strengths of ropes manufactured by the two mills are normally distributed with variances σ_1^2 and σ_2^2, respectively.

44. Repeat Problem 36, assuming that the two populations are normally distributed with unequal variances.

45. Referring to Problem 1 of Section 8.7, find a 99% confidence interval for p, the proportion of people in the population wearing glasses.

46. Referring to Problem 2 of Section 8.7, find 90% and 95% confidence intervals for p, the probability of obtaining a head when this coin is tossed.

47. Referring to Problem 3 of Section 8.7, find a 99% confidence interval for the difference in proportions of the population favoring Brand X before and after the advertising campaign. Comment on your result.

48. Repeat Problem 4 of Section 8.7, using 98% and 99% confidence levels.

49. Referring to Problem 5 of Section 8.7, find a 99% confidence interval for the difference between the percentages of persons who favor a nuclear plant in their state. Compare the confidence intervals you obtained here to the one you obtained in Problem 5 of Section 8.7 and comment on the effect of increasing the confidence level from 95% to 99%.

Chapter 9

HYPOTHESIS TESTING

> *. . . the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis.*
>
> R.A. Fisher

The focus of this chapter is a discussion of testing of statistical hypotheses.

Topics Covered

- Basic concepts of testing of statistical hypotheses
- Tests concerning the mean of a normal distribution when variance is known
- Tests concerning the mean of a normal distribution when variance is unknown
- Tests concerning population means when the sample size is large
- Tests concerning the difference of means of two populations with known variances
- Tests concerning the difference of means of two populations with unknown variances
- The paired t-test
- Testing concerning one and two population proportions when the sample size is large
- Tests concerning the variance of a normal distribution
- Tests concerning the ratio of variances of two normal populations
- Sequential tests of hypotheses

Learning Outcomes

After studying this chapter, the reader will be able to

- Construct null and alternative hypotheses.
- Determine an appropriate test statistic and use it to carry out a hypothesis test.
- Understand the concepts of type I and type II errors, and determine the power of a test.
- Understand the concept of the p-value, calculate it, and then use it to make the correct decision.
- Use appropriate confidence intervals to carry out various tests of hypotheses.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

9.1 INTRODUCTION

In our discussions in earlier chapters, we noted that one of the aims of statistics is to make inferences about the unknown parameters of a population, based on the information contained in a sample that is selected from this population. The goal of making such inferences may be achieved by estimating the unknown parameters and then by testing hypotheses about the plausible values of these unknown parameters. In Chapter 8, we considered the problem of estimating the unknown parameters. Here, we consider certain aspects of statistical testing of hypotheses.

Testing of hypotheses is a phenomenon that we deal with in everyday life. For example, a pharmaceutical company may like to test a certain hypothesis about a new drug used to treat patients with high cholesterol, breast cancer, or coronary artery disease. Amtrak, a train transportation service company, may like to test whether an existing track can be used to introduce a new train service for a particular route that covers a certain distance in a given period of time. A quality engineer in a paper mill may test a hypothesis that the new machine will produce no more than 10% of paper with defects. A civil engineer may like to test a hypothesis that a new bridge can withstand a weight of 80 tons. Even the United States Congress may test a hypothesis that the new economic measures can reduce the unemployment rate by one full point.

Another type of problem that may arise during hypothesis testing, concerns whether a sample could reasonably have come from a population having a completely or partially specified distribution. For instance, if a sample is known to have come from some normal distribution, is it reasonable that it could have come from one having a given mean μ_0? Or if two independent samples come from normal distributions, is it reasonable that they could have come from normal distributions with equal means?

To inquire about such hypotheses, we are obliged to collect some data, meaning draw samples from given populations and test the validity of these hypotheses utilizing the sample data. We then proceed by making use of sample averages, proportions, variances, and other statistics determined from the sample or samples. Statistics such as these, when determined from samples, are random variables having their own probability distributions, so statements based on their values must be made in terms of probabilities. In this chapter, we consider some of the more important statistical tests based on sample averages and sample variances that can help us either establish or contradict, with a certain desired probability, the validity of such hypotheses. Also, as will be seen, there is a close connection between statistical testing and statistical estimation.

9.2 BASIC CONCEPTS OF TESTING A STATISTICAL HYPOTHESIS

9.2.1 Hypothesis Formulation

The first step toward testing a statistical hypothesis is to identify an appropriate probability model for the population under investigation and to identify the parameter around which the hypothesis is being formulated. For example, if we identify a normal probability model as an appropriate model for the population under investigation, then we may formulate a hypothesis about the mean μ and/or the standard deviation σ. Once an appropriate probability model is selected and the hypothesis is formulated, then the subsequent steps are to collect data and conduct the testing of the hypothesis we had formulated, leading to deciding whether we support or discredit the hypothesis with a certain desirable probability.

Generally speaking, a statistical hypothesis consists of a pair of statements about the unknown parameter. One of these statements describes someone's belief or the existing theory, which is called the *null hypothesis* and denoted by H_0. The second statement is usually an assertion that is made based on some new information. It is called the *research hypothesis* or an *alternative hypothesis* and is denoted by H_1.

Then, from the information contained in a sample, we either reject the null hypothesis H_0 in favor of the alternative hypothesis H_1 or we do not reject H_0.

For example, suppose that we have a population with a probability model $f(x, \theta)$, where θ is an unknown parameter. Then, we can formulate a statistical hypothesis described as

$$H_0\colon \theta = \theta_0, \ H_1\colon \theta < \theta_0 \tag{9.2.1}$$

where θ_0 is known. Thus, under H_0, the null hypothesis, it is believed that θ takes a known value θ_0, whereas under H_1, the alternative hypothesis, our assertion, based on some new information, is that θ takes a value less than θ_0. Should we have some different information, then that could lead us to another alternative hypothesis, namely

$$H_1'\colon \theta > \theta_0 \ \text{ or } \ H_1''\colon \theta \neq \theta_0 \tag{9.2.2}$$

Note that under the null hypothesis H_0, we have a specified value θ_0 of θ, whereas under the alternative hypotheses, we do not have any specified value of θ. A hypothesis that assigns a specified value to an unknown parameter is called a *simple hypothesis* and one that does not assign a specified value to the unknown parameter is called a *composite hypothesis.* The alternative hypotheses

$$H_1\colon \theta < \theta_0 \ \text{ and/or } \ H_1'\colon \theta > \theta_0 \tag{9.2.3}$$

are called *one-sided* or *one-tail alternatives*, whereas

$$H_1''\colon \theta \neq \theta_0 \tag{9.2.4}$$

is called a *two-sided* or *two-tail alternative.*

With this terminology, we now describe a general procedure used to test these hypotheses. As we remarked earlier, to test a hypothesis, we use the information contained in a sample that has been drawn from the population with probability model $f(x, \theta)$, where θ is an *unknown parameter.* This is done by considering some statistic, called the *test statistic*, say $\hat{\theta}$, which may be an estimator of θ. Then using the sample data, we calculate the value of the test statistic. For certain values of the test statistic, we may favor the alternative hypothesis H_1 and reject the null hypothesis H_0, whereas for other values of the test statistic, the null hypothesis H_0 would not be rejected. For example, consider the following hypotheses:

$$H_0\colon \theta = \theta_0 \ \text{ versus } \ H_1\colon \theta < \theta_0$$

It seems reasonable to consider that if the value of the test statistic $\hat{\theta}$ turns out to be "too small", then we should favor the alternative hypothesis H_1 and reject the null hypothesis H_0. Otherwise, we should not reject H_0.

Deciding how small of a value of $\hat{\theta}$ is "too small" can be done by considering the sample space of $\hat{\theta}$ and dividing it into two regions so that if the value of $\hat{\theta}$ falls in the lower

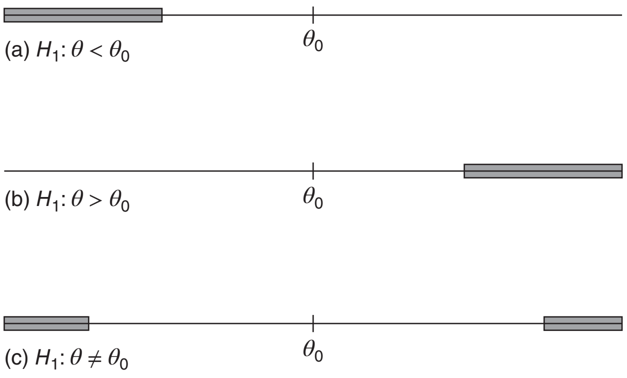

(a) $H_1: \theta < \theta_0$
θ_0
(b) $H_1: \theta > \theta_0$
θ_0
(c) $H_1: \theta \neq \theta_0$
θ_0

Figure 9.2.1 Critical points dividing the sample space of $\hat{\theta}$ in two regions, the rejection region (shaded) and the acceptance region (nonshaded).

region (the shaded region in Figure 9.2.1a, we reject H_0. Otherwise, we do not reject H_0. The region for which we reject H_0 is usually called the *rejection region* or *critical region*, and the region for which we do not reject the null hypothesis H_0 is called the *acceptance region*. The point separating these two regions is called the *critical point*. Using the same argument, we can easily see that for testing H_0 against the alternatives

$$H_1: \theta > \theta_0 \text{ or } H_1: \theta \neq \theta_0$$

the hypothesis $H_1: \theta > \theta_0$ is favored for large values of $\hat{\theta}$, while the hypothesis $H_1: \theta \neq \theta_0$ is favored when $\hat{\theta}$ is either very small or very large. Thus, the rejection regions will fall, respectively, in the upper region, and in both lower and upper regions, as summarized in Figure 9.2.1b and c, respectively.

9.2.2 Risk Assessment

We have now developed a procedure for using the information contained in a sample by means of a statistic, taking a decision about the unknown parameters, and consequently about the population itself. The next question that we might ask is whether there is any risk of committing any errors while making such decisions. The answer to this question is yes. There are two risks. The first occurs when the null hypothesis is true, but based on the information contained in the sample, we end up rejecting H_0. This type of error is called *type I error*. The second kind of error occurs when the null hypothesis is false; that is, the alternative hypothesis is true but still we do not reject the null hypothesis. This kind of error is called *type II error*. These errors cannot be eliminated completely, but they can certainly be minimized by taking large samples. We study this aspect of the problem later in this chapter.

There are certain probabilities associated with committing the type I and type II errors that we denote by α and β, respectively. Thus, we may define α and β as follows:

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true}) \tag{9.2.5}$$

$$\beta = P(\text{do not reject } H_0 | H_0 \text{ is false}) \tag{9.2.6}$$

Table 9.2.1 Type I and type II errors and their probabilities of occurrence (in parentheses).

	H_0 is true	H_0 is false
Reject H_0	Type I error (α)	Correct decision
Do not reject H_0	Correct decision	Type II error (β)

At this point, it is useful to summarize in Table 9.2.1 the discussion so far of type I and type II errors and their probabilities.

Some terminology connected to the table above is as follows:

1. $\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) =$ *level of significance* = size of the test. In quality control, α is commonly known as the *producer's risk*.
2. $\beta = P(\text{type II error}) = P(\text{do not reject } H_0 | H_0 \text{ is false})$. In quality control, β is commonly known as the *consumer's risk*.
3. We note that the complement of β is called the *power of the tes*t, that is, $Power = 1 - \beta = 1 - P(\text{do not reject } H_0 | H_0 \text{ is false}) = P(\text{reject } H_0 | H_0 \text{ is false})$ The power of a test is often denoted by γ, that is, $\gamma = 1 - \beta$.

The value of β depends on the alternative hypothesis, and β is always determined at a specific value of θ under the alternative hypothesis. For example, if the population is $N(\mu, \sigma^2)$, and if we are considering the null and alternative hypotheses $H_0\colon \mu = \mu_0$ and $H_1\colon \mu < \mu_0$, respectively, then a specific value of the alternative, say μ_1, should be such that μ_1 is less than μ_0. Similarly, if the alternative hypothesis is $H_1\colon \mu > \mu_0$ or $H_1\colon \mu \neq \mu_0$, then μ_1 should be such that $\mu_1 > \mu_0$ or $\mu \neq \mu_0$, respectively.

Now in *setting up a test of a hypothesis* H_0, when the alternative is H_1, there are some useful steps to follow. These are as follows:

Step 1. State the null hypothesis and the alternative hypothesis very clearly.

Step 2. Assign an appropriate value to the level of significance, that is, α. It is very common in practice to assign 0.01, or 0.05, or 0.10 for the value of α.

Step 3. Determine a suitable test statistic. For the statistical hypotheses that we are going to discuss in this and other chapters, the pivotal quantity under H_0 (see Chapter 8) for the parameter under investigation is often used as a test statistic.

Step 4. Determine the probability distribution of the test statistic designated in Step 3.

Step 5. Locate the rejection region(s) and determine the critical point. The location of the rejection region always depends on the alternative hypothesis while the size of the rejection region depends on the value assigned to α, the probability of the type I error.

Step 6. Calculate the value of the test statistic and make the decision. That is, take a random sample from the population in question and calculate the value of the test statistic. Then verify whether or not the value of the test statistic falls in the rejection region. If it falls in the rejection region, then we reject the null hypothesis H_0. Otherwise, we do not reject H_0.

The reader may notice that the value of α is prechosen by the experimenter. However, the value of β often needs to be determined, which sometimes may become cumbersome. If the probability distribution of the test statistic (pivotal quantity under H_0) is normal, for example, $\bar{X} \sim N(\mu, \sigma^2/n)$, then the value of β can easily be obtained by using one of the appropriate formulas given below (σ assumed known); see Section 9.3 for details.

$$\beta = P\left(Z > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_\alpha\right) = 1 - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_\alpha\right) \quad \text{if } H_1\colon \mu_1 < \mu_0 \qquad (9.2.7)$$

$$\beta = P\left(Z < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha\right) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha\right) \quad \text{if } H_1\colon \mu_1 > \mu_0 \qquad (9.2.8)$$

$$\begin{aligned}\beta &= P\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2} < Z < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right)\\ &= \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) \quad \text{if } H_1\colon \mu_1 \neq \mu_0 \qquad (9.2.9)\end{aligned}$$

Clearly, when the value of β decreases, the power of the test increases, and therefore, we have a better test. Thus, at this juncture, one might ask whether there are any circumstances under which one can assign some predetermined value to β as well. The answer to this question is yes, and it can be done by selecting an appropriate sample size that may turn out to be quite large. For the case σ known, given values of α and β, the sample size n should be such that

$$n \geq \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu_1 - \mu_0)^2} \quad \text{for a one-tail test} \qquad (9.2.10)$$

$$n \geq \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{(\mu_1 - \mu_0)^2} \quad \text{for a two-tail test} \qquad (9.2.11)$$

We will turn to a discussion of various concrete situations, after the practice problems for this section.

PRACTICE PROBLEMS FOR SECTION 9.2

1. A random sample of $n = 36$ observations from a population with mean μ and variance σ^2, where $\sigma^2 = 0.25$, produced a sample mean $\bar{X} = 4.2$ and a sample standard deviation equal to 0.5. If it is desired to test the hypothesis that the population mean

μ exceeds 4.0, describe the null hypothesis and alternative hypothesis and carry out the testing of hypothesis at the 5% level of significance.

2. Referring to Problem 1, suppose that the type I error is 0.05 and the true population mean is 4.5. Determine the size of β the probability of the type II error.
3. Referring to Problem 2, if the true population mean is 5.0 instead of 4.5, what is your conjecture about the size of the probability of the type II error (β) of the test; that is; will it be larger or smaller? Determine the actual value of β and check whether or not your conjecture was right.
4. In a given hypothesis testing problem, how is the value of the probability of the type II error (β) affected if the sample size remains the same but the value of the type I error (α) changes?
5. Determine how large a sample size should be taken if the test for Problem 1 is such that $\alpha = 0.05, \beta = 0.10, \mu_0 = 4.0$, and $\mu_1 = 4.2$.

9.3 TESTS CONCERNING THE MEAN OF A NORMAL POPULATION HAVING KNOWN VARIANCE

9.3.1 Case of a One-Tail (Left-Sided) Test

Suppose that $\bar{X}$ is the average of a sample of size n from a normal distribution $N(\mu, \sigma^2)$, where σ^2 is known and μ is unknown. Suppose that we want to learn whether it is reasonable to think that this sample came from the normal population $N(\mu_0, \sigma^2)$ as compared with the possibility it came from another normal population $N(\mu_1, \sigma^2)$, where $\mu_1 < \mu_0$. We can abbreviate this statement to test this statistical hypothesis (or null hypothesis) H_0 versus the alternative H_1, where H_0 and H_1 are such that

$$H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu < \mu_0$$

and where we are making use of a sample of size n that has the sample mean $\bar{X}$, which is an estimator of μ.

It is intuitively evident that we would choose H_1 if $\bar{X}$ is sufficiently small, that is, if $\bar{X} < k$, where k is yet to be found, and favor H_0 if $\bar{X} \geq k$. The set of values of $\bar{X}$ for which we reject H_0 (i.e., those for which $\bar{X} < k$) is called the *critical region* for the test. In making any decision, we could make two kinds of errors (see Table 9.2.1).

However, we always test a hypothesis under the assumption that the null hypothesis is true, and we must either support this assumption or contradict it. That is, if we support our assumption, then we do not reject H_0. Otherwise, we reject H_0. Now by choosing k so that

$$P(\bar{X} < k | \mu = \mu_0) = \alpha \tag{9.3.1}$$

we can control the type I error so that its probability of occurrence is α. The type II error has probability β, where for $\mu = \mu_1 < \mu_0$,

$$\beta = P(\bar{X} \geq k | \mu = \mu_1) = 1 - P(\bar{X} < k | \mu = \mu_1) \tag{9.3.2}$$

In this case, k has been chosen to satisfy equation (9.3.1).

Since α is known and k has been chosen so that equation (9.3.1) is satisfied and since $\bar{X}$ has the distribution $N(\mu_0, \sigma^2/n)$ in equation (9.3.1), it is seen that

$$P(\bar{X} < k|\mu = \mu_0) = \Phi\left[\frac{(k-\mu_0)\sqrt{n}}{\sigma}\right] = \alpha \tag{9.3.3}$$

Recalling equation (9.3.2), we have

$$\beta = P(\bar{X} \geq k|\mu = \mu_1) = 1 - \Phi\left(\frac{k-\mu_1}{\sigma/\sqrt{n}}\right) \tag{9.3.4}$$

Thus, for specified values of $\alpha, \mu_0, \mu_1, \sigma$, and n, from equation (9.3.3) it follows that the number k satisfies the equation

$$\frac{(k-\mu_0)\sqrt{n}}{\sigma} = z_{1-\alpha} = -z_\alpha \tag{9.3.5}$$

Solving equation (9.3.5), we find

$$k = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} = \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha \tag{9.3.6}$$

and hence the critical region for $\bar{X}$ is the set of values of $\bar{X}$ for which $\bar{X} < k$, that is,

$$\bar{X} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} = \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha \tag{9.3.7}$$

or for which

$$\frac{(\bar{X}-\mu_0)\sqrt{n}}{\sigma} < z_{1-\alpha} = -z_\alpha \tag{9.3.8}$$

By substituting the value of k from equation (9.3.6) in equation (9.3.4), we find that the probability of committing a type II error when $\mu = \mu_1 < \mu_0$ is

$$\beta = P\ (\text{Do not reject } H_0|H_0 \text{ is false})$$

so that

$$\begin{aligned}
\beta &= P(\bar{X} \geq \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha|\mu = \mu_1) \\
&= P(\bar{X} - \mu_1 \geq (\mu_0 - \mu_1) - \frac{\sigma}{\sqrt{n}} z_\alpha|\mu = \mu_1) \\
&= P\left(\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}} \geq \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_\alpha|\mu = \mu_1\right) \\
&= P\left(Z \geq \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_\alpha\right)
\end{aligned} \tag{9.3.9}$$

We note that any observed value of $\bar{X}$ falling in the critical region in equation (9.3.7) is said to be *significantly smaller than* μ_0 at the $100\alpha\%$ level of significance.

Now, as noted in equation (9.3.8), the critical region in equation (9.3.7) can also be expressed as the set of values of $\bar{X}$ for which

$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha \tag{9.3.10}$$

so that the description of the critical region uses

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{9.3.11}$$

the pivotal quantity (see Chapter 8) for μ evaluated under H_0, that is, evaluated at $\mu = \mu_0$. Now under H_0: $\mu = \mu_0$, the test statistic in equation (9.3.11) has the distribution of the standard normal random variable Z, so that the size of the test, the probability of committing the type I error of the test H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$ is

$$P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha | \mu = \mu_0\right) = P\left(Z < -z_\alpha\right) = \alpha$$

as stated previously. The critical region of size $\alpha = 0.05$ falls under the left tail of the standard normal distribution (see Figure 9.3.1).

Note that β as given by equation (9.3.9) is a function of μ_1, say $\beta(\mu_1)$, and the graph of $\beta(\mu_1)$, the probability of type II error plotted against μ_1 is called the *operating characteristic* (*OC*) *curve* of the test of H_0 against H_1 (see Figure 9.3.2). The function

$$\gamma(\mu_1) = 1 - \beta(\mu_1) \tag{9.3.11a}$$

where $\mu_1 < \mu_0$, is called the power function of the test and gives the probability of rejecting $\mu = \mu_0$, given that $\mu = \mu_1$. The graph of the power function plotted against μ_1 is called the power curve of the test. Note that the ordinate of the power curve at μ_0 is α, that is,

$$\gamma(\mu_0) = 1 - \beta(\mu_0) = \alpha$$

Often the experimenter wishes to cite the observed level of significance, often called the p-value where, for example, in this situation

$$p\text{-value} = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x}_{obs} - \mu_0}{\sigma/\sqrt{n}} | \mu = \mu_0\right) = P\left(Z < \frac{\bar{x}_{obs} - \mu_0}{\sigma/\sqrt{n}}\right) \tag{9.3.12}$$

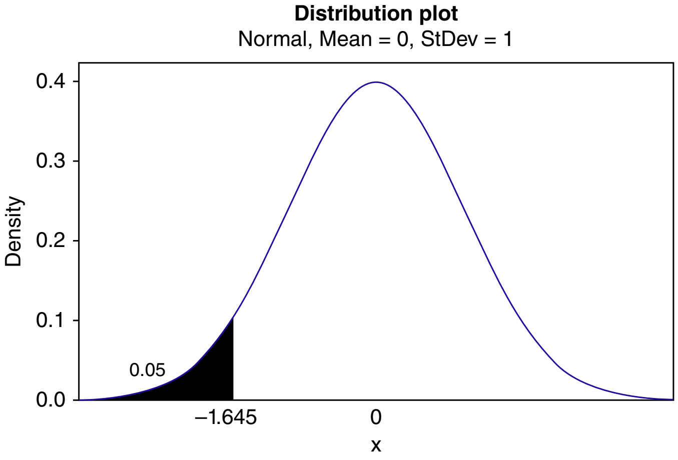


Figure 9.3.1 The critical region for the left-sided test with $\alpha = 0.05$.

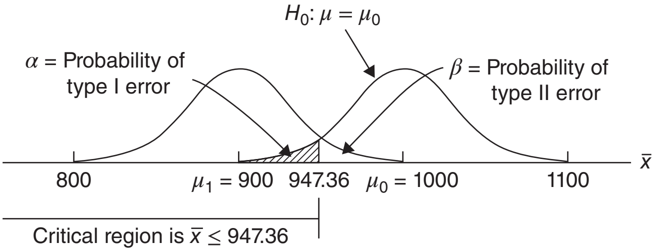

H_0: $\mu = \mu_0$
α = Probability of type I error
β = Probability of type II error
800
μ_1 = 900
947.36
μ_0 = 1000
1100
$\bar{x}$
Critical region is $\bar{x} \leq 947.36$

Figure 9.3.2 Graphical description of the constituents of Example 9.3.1.

where $\bar{x}_{obs}$ is the observed value of the sample average $\bar{X}$. In general, we may define the p-value as follows:

Definition 9.3.1 The p-value of a test is the smallest value of α for which the null hypothesis H_0 is rejected.

Suppose that the *observed value* of the test statistic Z is z, where $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$. Then the p-value for observing $z = (\bar{x}_{obs} - \mu_0)/(\sigma/\sqrt{n})$, is given by (see equation (9.3.12))

$$\begin{aligned} p\text{-value} &= P(Z \leq z) && if\ H_1\colon \mu < \mu_0 \\ &= P(Z \geq z) && if\ H_1\colon \mu > \mu_0 \\ &= 2P(Z \geq |z|) && if\ H_1\colon \mu \neq \mu_0 \end{aligned} \tag{9.3.12a}$$

The rules for using the p-value defined previously are the following:

1. If p-value $< \alpha$, we reject H_0.
2. If p-value $\geq \alpha$, we do not reject H_0.

We illustrate with the following example.

Example 9.3.1 (Testing a left-sided hypothesis about μ when σ is known) *A sample of 16 lengths of wire from a day's production on a given machine has an average tensile strength* $\bar{X} = 967.8$ *psi. Suppose that the population of tensile strengths of wire in the day's production is* $N(\mu, (128)^2)$ *(it is known from experience that for this type of wire* $\sigma = 128$ *psi) and that we wish to test the hypothesis*

$$H_0\colon \mu = 1000 \textit{ versus } H_1\colon \mu < 1000$$

In this case, we have from equation (9.3.7) that

$$k = 1000 - (128/\sqrt{16})z_\alpha$$

If we choose α, the probability of a type I error, to be 0.05, then $z_\alpha = z_{0.05} = 1.645$, then

$$k = 1000 - 32(1.645) = 947.36$$

and the critical region consists of the values of $\bar{X}$ for which $\bar{X} < 947.36$. Since the observed value of $\bar{X}$ is 967.8, which does not fall in the critical region $(-\infty, 947.36)$, we say that $\bar{X}$ is not significantly smaller than the value of μ under H_0, or that, at the given $\alpha = 0.05$ level, we do not reject the null hypothesis $\mu = 1000$. We observe here that the p-value is $P\left(Z < (967.8 - 1000)\sqrt{16}/128\right) = 0.1562$, which is greater than $\alpha = 0.05$, confirming the above decision that H_0 is not rejected.

The probability of the type II error $\beta(\mu_1)$ of our test for H_0 against any $\mu_1 < 1000$ is the probability of not rejecting H_0 (i.e., not rejecting the hypothesis that $\mu = 1000$) when H_1 is true (i.e., when μ has some value $\mu_1 < 1000$). Suppose that now $\mu_1 = 900$. Then we use equation (9.3.9), i.e., $\beta(\mu_1)$, to obtain the probability of type II error as

$$\begin{aligned}\beta(\mu_1 = 900) &= P\left(Z \geq \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} - z_\alpha\right) = P\left(Z \geq \frac{1000 - 900}{128/\sqrt{16}} - 1.645\right)\\ &= P(Z \geq 1.48) = 0.0694\end{aligned}$$

and the power of the test at $\mu_1 = 900$ is given by (see equation (9.3.11a))

$$\gamma(\mu_1 = 900) = 1 - 0.0694 = 0.9306$$

Figure 9.3.2 depicts α and β, the value of k, $\mu_0(\mu_0 = 1000)$, and $\mu_1(\mu_1 = 900)$.

9.3.2 Case of a One-Tail (Right-Sided) Test

Suppose that we want to test, at significance level α, the hypothesis testing problem

$$H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu > \mu_0$$

on the basis of the average $\bar{X}$ of a sample of size n from $N(\mu, \sigma^2)$, where σ^2 is known. The reader can verify that in this case, the critical region for $\bar{X}$ is the set of values of for which $\bar{X} > k$, where k satisfies

$$P(\bar{X} > k|\mu = \mu_0) = \alpha \tag{9.3.13}$$

We easily find that k is such that

$$k = \mu_0 + \frac{\sigma}{\sqrt{n}}z_\alpha \tag{9.3.13a}$$

so that we reject $H_0\colon \mu = \mu_0$ when testing against the alternative $H_1\colon \mu > \mu_0$ if

$$\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}z_\alpha \tag{9.3.14}$$

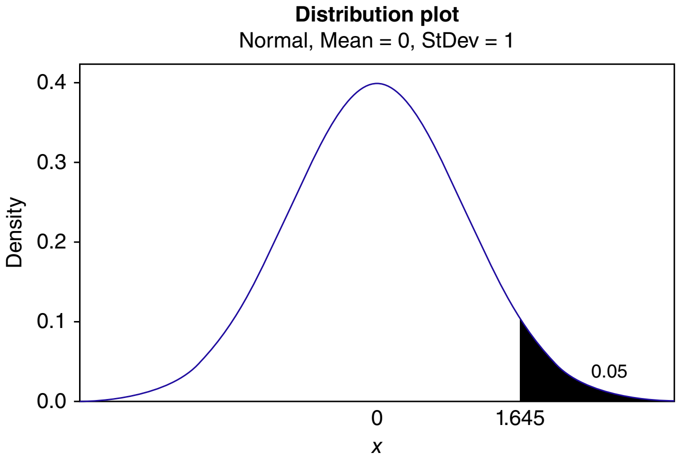


Figure 9.3.3 Critical region for the right-sided test with $\alpha = 0.05$.

The probability β of a type II error for this test is easily seen to be (see equation (9.2.8))

$$\beta = P\left(Z < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha\right) \tag{9.3.15}$$

where $\mu_1 > \mu_0$.

We note that since the critical region consists of values of $\bar{X}$, which are such that

$$\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \tag{9.3.16}$$

we often say that we reject H_0 at significance level α if

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \tag{9.3.17}$$

Again, we use the pivotal quantity evaluated at $\mu = \mu_0$, namely

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{9.3.18}$$

as a test statistic and its critical region of size α falls under the right tail of the standard normal distribution (see Figure 9.3.3).

9.3.3 Case of a Two-Tail Test

Now suppose that we want to test the hypothesis

$$H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu \neq \mu_0$$

In this case, it is evident that the critical region of the test will consist of all values of $\bar{X}$ for which $|\bar{X} - \mu_0| > k$ where, for a given probability α of a type I error, k is chosen so that

$$P(|\bar{X} - \mu_0| > k | \mu = \mu_0) = \alpha$$

That is,

$$1 - P\left(-\frac{k}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{k}{\sigma/\sqrt{n}} | \mu = \mu_0\right) = \alpha \tag{9.3.19}$$

Hence, we must have

$$1 - \left(\Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{k}{\sigma/\sqrt{n}}\right)\right) = \alpha \tag{9.3.20}$$

where Φ is the cumulative distribution function (c.d.f.) of $Z, Z \sim N(0,1)$. But remembering that $\Phi(z) + \Phi(-z) = 1$, we have from equation (9.3.20), after some simplification, that

$$\Phi\left(-\frac{k}{\sigma/\sqrt{n}}\right) = \frac{\alpha}{2}$$

Hence, we have

$$-\frac{k}{\sigma/\sqrt{n}} = -z_{\alpha/2} \text{ or } \frac{k}{\sigma/\sqrt{n}} = z_{\alpha/2}$$

The critical region for $\bar{X}$ is thus the set of values for $\bar{X}$ for which

$$\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2} \tag{9.3.21}$$

If $\bar{X}$ satisfies equation (9.3.21), we say that $\bar{X}$ *differs from* μ_0 *significantly at the* α *level of significance* and we reject H_0.

The critical region given in equation (9.3.21) with $\alpha = 0.05$ can be depicted graphically under the standard normal frequency curve as shown in Figure 9.3.4. The power of two-sided test for H_0 against H_1 is illustrated in Example 9.3.2.

The reader should also note that since

$$P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2} | \mu = \mu_0\right) = \alpha$$

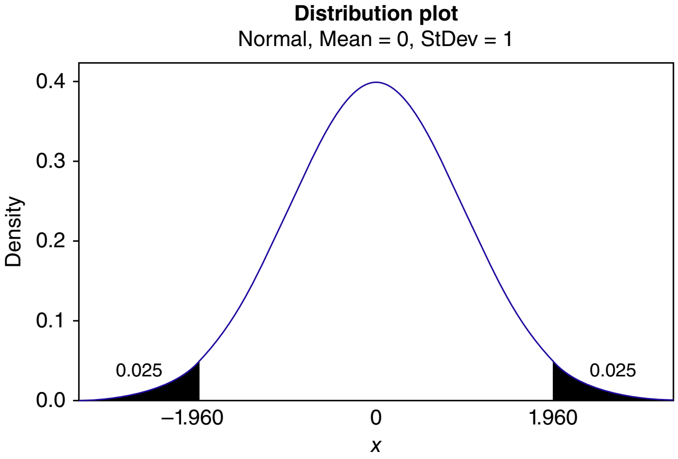


Figure 9.3.4 Critical region for the two-sided test with $\alpha = 0.05$.

we have

$$P\left(|\bar{X} - \mu_0| < \frac{\sigma}{\sqrt{n}} z_{\alpha/2} | \mu = \mu_0\right) = 1 - \alpha$$

which can be rewritten as

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu_0 < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1 - \alpha$$

This result is equivalent to the statement that $[\bar{X} \pm (\sigma/\sqrt{n}) z_{\alpha/2}]$, the $100(1-\alpha)\%$ confidence interval for μ, contains μ_0 the value specified by H_0: $\mu = \mu_0$.

We summarize this result as follows:

$\bar{X}$ being (not being) significantly different from μ_0 at the α level of significance is equivalent to the statement that μ_0 does not (does) lie in the $100(1-\alpha)\%$ confidence interval for μ.

Example 9.3.2 (Testing a two-sided hypothesis about μ when σ is known) *Referring to Example 9.3.1, suppose that we test the hypothesis*

$$H_0\colon \mu = 1000 \text{ versus } H_1\colon \mu \neq 1000$$

at the 5% level of significance.

Solution: We have, for $\alpha = 0.05$, that $z_{\alpha/2} = z_{0.025} = 1.96$; hence from equation (9.3.21), the critical region for $\bar{X}$ is the set of values of $\bar{X}$ for which

$$|\bar{X} - 1000| > \frac{128}{\sqrt{16}} \times 1.96 \Rightarrow |\bar{X} - 1000| > 62.72$$

Hence, we reject H_0 if either $\bar{X} < 1000 - 62.72 = 937.28$ or $\bar{X} > 1000 + 62.72 = 1062.72$.

Here, the observed value of $\bar{X} = 967.8$ does not fall into this critical region. Thus, the observed value of $\bar{X}$, 967.8, does not differ significantly from the value $\mu_0 = 1000$ (specified under H_0) at the 5% level of significance, and we do not reject H_0.

Of course, we can reject H_0 or not reject H_0, on the basis of the value of the observed level of significance, namely the p-value for the test. To calculate the p-value, we first recall that the form of the test of H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$ at significance level α is that we reject H_0: $\mu = \mu_0$ if $\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2} = z_{0.025} = 1.96$. Hence, the p-value is given by

$$\begin{aligned} p\text{-value} &= P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x}_{obs} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \\ &= P\left(|Z| > \left|\frac{967.8 - 1000}{128/\sqrt{16}}\right|\right) \\ &= P(|Z| > 1.01) = 0.3124 \end{aligned}$$

Since we have that p-value > 0.05, we do not reject H_0.

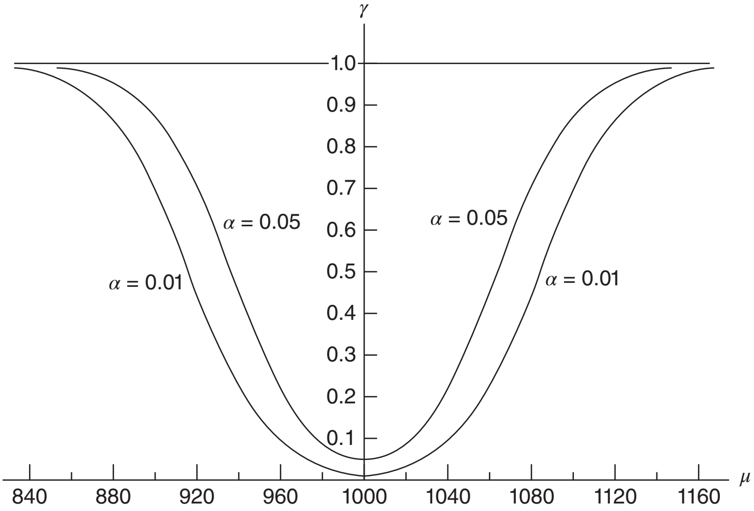


Figure 9.3.5 Power curves for the test in Example 9.3.2.

Notice that from the point of view of confidence intervals, we find that the value of $\mu, \mu = \mu_0 = 1000$ (under H_0) is contained in the 95% confidence interval for μ, namely (967.8 ± 62.72), so we do not reject H_0: $\mu = 1000$.

The power of this test (see equation (9.2.9)) with $\alpha = 0.05$ is easily seen. For $\mu = \mu_1 \neq \mu_0$, it is given by

$$\gamma(\mu_1) = P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > 1.96|\mu = \mu_1\right) = P(\text{Reject } H_0|H_1 \text{ is true})$$

But under H_1: $\mu = \mu_1$, $\dfrac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \sim Z$, $Z \sim N(0, 1)$, so we find

$$\gamma(\mu_1) = P(Z < (\mu_0 - \mu_1)/(\sigma/\sqrt{n}) - 1.96) + P(Z > (\mu_0 - \mu_1)/(\sigma/\sqrt{n}) + 1.96)$$

for $\mu_1 \neq \mu_0$. We plot $\gamma(\mu_1)$ in Figure 9.3.5. This plot also gives the power curve for a two-sided test of H_0: $\mu = 1000$ versus H_1: $\mu \neq 1000$ when $\alpha = 0.01$.

From Figure 9.3.5, it is clear that as the value of α increases, the power of the test also increases, which in turn implies that β the probability of type II error, decreases. Note that as long as the sample size and other conditions remain the same, α and β move in opposite directions, so that as α increases β decreases, and as α decreases β increases. However, the reader should also note that if the sample size increases, then both α and β decrease. It is also to be noted that sometimes, the question of *statistical significance* versus *practical significance* arises; that is, are our results statistically significant but without practical implications? We illustrate this point with an example.

Example 9.3.3 (Testing a left-sided hypothesis about μ when σ is known) *A random sample of 36 pieces of copper wire produced in a plant of a wire-manufacturing company yields the mean tensile strength of* $\bar{X} = 975$ *psi. Suppose that the population of tensile* *strengths of all copper wires produced in that plant is distributed with mean μ and standard deviation $\sigma = 120$ psi. Test at the $\alpha = 0.01$ level of significance the hypothesis*

$$H_0\colon \mu = 980 \text{ versus } H_1\colon \mu < 980$$

Solution: We solve this problem by taking the stepwise approach:

1. $H_0\colon \mu = 980$ versus $H_1\colon \mu < 980$
2. $\alpha = 0.01$
3. Use the test statistic
$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$
4. Since the sample size $n = 36$ (≥ 30) is large, then using the central limit theorem, the test statistic is assumed to be distributed as $N(0, 1)$ to good approximation.
5. Since the test is a left-sided test, use the standard normal distribution tables to show that the rejection region is as presented in Figure 9.3.1; that is, reject H_0 at significance level $\alpha = 0.01$ if
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_{0.01} = -2.326$$
6. The observed value of the test statistic is
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{975 - 980}{120/\sqrt{36}} = -0.25$$
which does not fall in the rejection region. Hence, the null hypothesis H_0 is not rejected.
7. The p-value is
$$P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -0.25|\mu = \mu_0 = 980\right) = P(Z < -0.25) = 0.4013$$
which is greater than $\alpha = 0.01$.

Thus, the data seem to support the hypothesis that the mean tensile strength of the copper wires manufactured in that plant is 980 psi. In other words, statistically, the observed value 975 psi is not significantly different from $\mu_0 = 980$ psi, and we say that the difference between the observed value (975 psi) and the desired value (980 psi) is *practically insignificant.*

Suppose now that we take a very large sample, say of 10,000 observations, and suppose that the sample average, again, comes out to be $\bar{X} = 975$ psi. The value of the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{975 - 980}{120/\sqrt{10{,}000}} = -4.16$$

and the p-value is $P(Z < -4.16) \approx 0$

Hence, we would reject the null hypothesis. In other words, statistically speaking, the observed value of 975 psi has become significantly different from the hypothesized value, 980 psi. Thus, the *statistical significance* has departed from the *practical significance.*

Example 9.3.4 (Using MINITAB and R) *The workers union of a large corporation in a big metropolitan city demands that each worker should be compensated for travel time to work since it takes, on the average, at least 75 minutes for each worker to travel to his/her job. However, the director of human resources believes otherwise and he took a random sample of 16 workers and found that the average traveling time for these workers is 68 minutes. Assume that from past experience, the director knows that travel times are normally distributed (in applications this condition must be verified) with a standard deviation $\sigma = 10$ minutes. Do these data provide sufficient evidence to support the directors's claim? Use $\alpha = 0.05$. Find the p-value. Find the size of the probability of the type II error β if the true travel time is $\mu = 72$ minutes.*

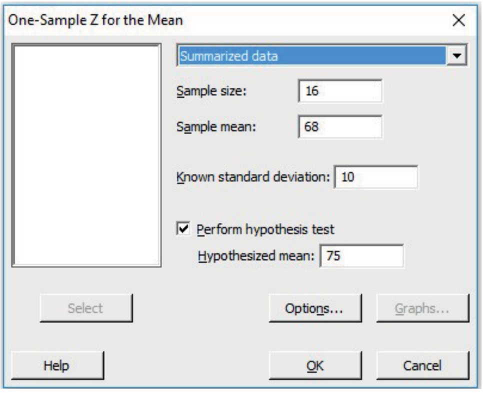
MINITAB

Based on the information given to us, the director would like to test the hypothesis

$$H_0\colon \mu = 75 \text{ versus } H_1\colon \mu < 75$$

To test a hypothesis for the mean at a given significance level using MINITAB, we proceed as follows:

1. Enter the data in column C1. If the summary statistics are given, as in this example, then skip this step.
2. If the population standard deviation is not known and the sample data is available, then using MINITAB calculate the sample standard deviation for the data.
3. Select **Stat** > **Basic Statistics** > **1-Sample Z** in the pull-down menu. This prompts a dialog box **One-Sample Z for the Mean** to appear on the screen. Note that we use **One-Sample Z** since the population is normal. If population is not given to be normal, then to use **1-Sample Z** command, the sample size must be larger (at least 30).
4. Select **Summarized data** from the pull-down menu in the dialog box and make the necessary entries as in the left panel shown below. If data are given, then select **One or more samples, each in a column** from the pulldown menu, and enter C1 in the box that appears below the pulldown menu (as in the right panel shown below). Enter the value of the population standard deviation in a box next to **Known standard deviation**.
5. Check the box next to **perform hypothesis test** and enter value of μ under the null hypothesis.


One-Sample Z for the Mean

Summarized data

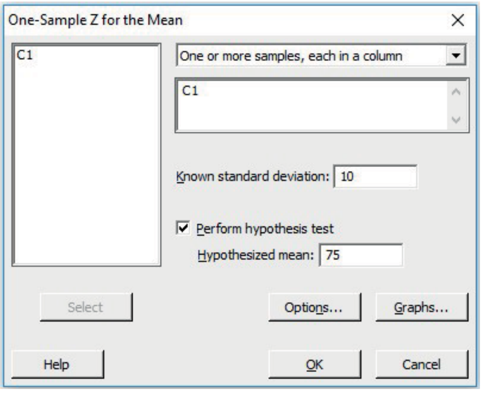
Sample size: 16

Sample mean: 68

Known standard deviation: 10

☑ Perform hypothesis test

Hypothesized mean: 75


One-Sample Z for the Mean

C1

One or more samples, each in a column

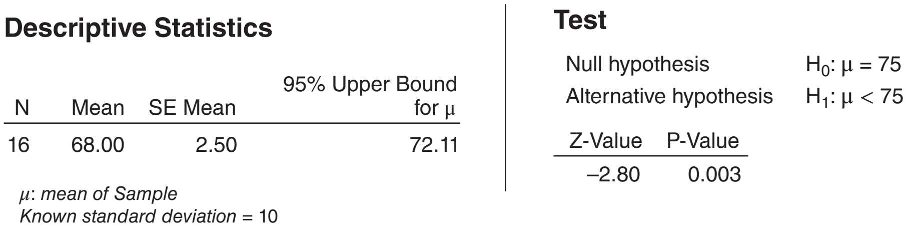
C1

Known standard deviation: 10

☑ Perform hypothesis test

Hypothesized mean: 75

6. Check **Options**, which prompts another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level** and enter the appropriate alternative hypothesis (not equal to, less than, or greater than). In each dialog box click **OK**. The MINITAB output shows up in the session window as given below:



```
Descriptive Statistics                              Test

                          95% Upper Bound           Null hypothesis         H0: μ = 75
N   Mean   SE Mean                  for μ           Alternative hypothesis  H1: μ < 75
-------------------------------------------
16  68.00     2.50                  72.11            Z-Value  P-Value
                                                    ------------------
 μ: mean of Sample                                    –2.80    0.003
 Known standard deviation = 10
```

Since the p-value is 0.003, which is less than the level of significance 0.05, we reject H_0 in favor of the alternative H_1, that is, the data do support the directors's claim.

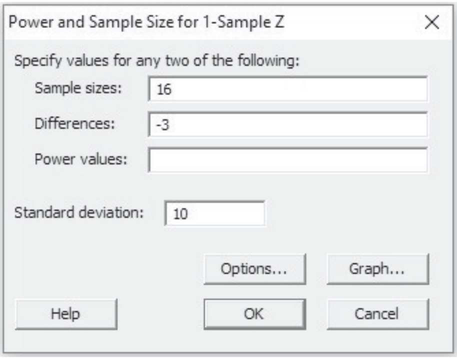
Interpretation: Based on this statistical testing procedure with significance level of 0.05, we do have sufficient evidence (p-value $= 0.003$) to conclude that on average it takes less than 75 minutes to travel to work place for its workers.

MINITAB is also equipped to find sample size and power of the test as long as we are given the value of one of them. The probability of the type II error β, at a given value of μ_1, is equal to $1 -$ power. In this example, we are given the sample size, so we proceed to find the power of the test, if $\mu_1 = 72$, as follows:

1. Select **Stat** > **Basic Statistics** > **1-Sample Z** (or one of the other options such as **1-Sample t, 2-Sample t**, or other appropriate choice). This prompts a dialog box **Power and Sample Size for 1-Sample Z** to appear on the screen.
2. In this dialog box, enter the value of **Sample size** (or **power value**) if you want to find the value of power (or sample size) and the **Difference**, where

$$\text{Difference} = \mu_1 - \mu_0 = 72 - 75 = -3$$

3. Enter the value of **Standard Deviation**.



Power and Sample Size for 1-Sample Z

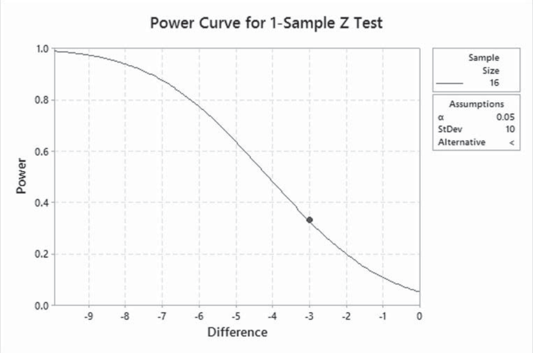
Specify values for any two of the following:

Sample sizes: 16

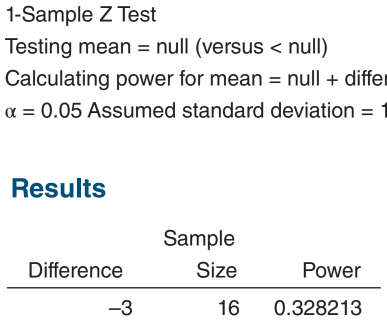
Differences: -3

Power values:

Standard deviation: 10



4. Select Options command and make the necessary entries in the new dialog box. In each dialog box, click **OK**. The MINITAB output shows up in the session window as given below, and it provides the above power curve.



```
1-Sample Z Test
Testing mean = null (versus < null)
Calculating power for mean = null + difference
α = 0.05 Assumed standard deviation = 10

Results

                  Sample
Difference          Size      Power
-----------------------------------
        –3            16   0.328213
```

We have $\gamma(72) = 0.3282$ so that the type II error $\beta = 1 - 0.3282 = 0.6718$.

USING R

Since, in this example, only the summary statistics were provided, we can use the built in ‘zsum.test()’ function in the ‘BSDA’ library in R. To conduct the test in Example 9.3.4, we may run the following code in the R Console window:

```
install.packages("BSDA")
library(BSDA)
zsum.test(mean.x = 68, sigma.x = 10, n.x = 16, alternative = "less",
          mu = 75, conf.level = 0.95)

#R output
One-sample z-Test
data: Summarized x
z = -2.8, p-value = 0.002555
alternative hypothesis: true mean is less than 75
```

For these data, we find that the value of the Z test statistic is -2.8, and the calculations return the corresponding p-value of 0.003. Since the p-value is less than the alpha-level of 0.05, we reject the null hypothesis as we did in MINITAB.

To determine the type II error, first we calculate the power at $\mu = 72$. This can be done by using the ‘pwr.norm.test()’ function in ‘pwr’ library in R. The following R code can be used to complete Example 9.3.4. Note that the value of effect size d is calculated as d $= (\mu_1 - \mu_0)/\sigma = (72 - 75)/10 = -0.3$.

```
install.packages("pwr")
library(pwr)
pwr.norm.test(d = -0.3, n = 16, sig.level = 0.05, alternative = "less")

#R output
Mean power calculation for normal distribution with known variance
d = -0.3
n = 16
sig.level = 0.05
power = 0.3282128
alternative = less
```

We have $\gamma(72) = 0.3282$ so that the type II error $\beta = 1 - 0.3282 = 0.6717872$.

PRACTICE PROBLEMS FOR SECTION 9.3

1. An existing process used to manufacture paint yields daily batches that have been fairly well established to be normally distributed with mean $\mu = 800$ tons, $\sigma = 30$ tons. A modification of this process is suggested with the view of increasing production. Assume that the daily yields, using the modified process, are distributed as $N(\mu, (30)^2)$, and suppose that a sample taken on 100 randomly chosen days of production using the modified process yields an average of $\bar{X} = 812$ tons. Test at the 1% level of significance H_0: $\mu = 800$ versus H_1: $\mu > 800$. What is the power of the test at $\mu = 810$? Graph the power function.
2. A machine used for producing "quarter inch rivets is to be checked by taking a random sample of 10 rivets and measuring their diameters. It is feared that the wear-off factor of the machine will cause it to produce rivets with diameters less than 1/4 in. Describe the critical region in terms of $\bar{X}$, the average of the 10 diameters, for a 1% significance test of H_0: $\mu = 0.25$ versus H_1: $\mu < 0.25$. Assume that the diameters are distributed as $N(\mu, (0.0015)^2)$ for a wide range of values of μ. What is the power of the test at $\mu = 0.2490$? Graph the power curve of the test.
3. Referring to Problem 1, suppose that the sample taken on 100 randomly chosen days of production using the modified process yields an average of $\bar{X} = 785$ tons. Test at the 1% level of significance: H_0: $\mu = 800$ versus H_1: $\mu < 800$. What is the power of the test at $\mu = 790$?
4. Referring to Problem 1, test at the 5% level of significance: H_0: $\mu = 800$ versus H_1: $\mu \neq 800$. What is the power of the test at $\mu = 795$ and at $\mu = 805$?
5. Refer to Problem 2. Describe the critical region in terms of $\bar{X}$, the average of the 10 diameters, for a test at level of significance of 1%, of H_0: $\mu = 0.25$ versus H_1: $\mu \neq 0.25$. What is the power of the test at $\mu = 0.2490$?
6. Referring to Problem 2, suppose a new machine was installed recently, and a random sample of 25 rivets produced yielded an average of the diameters of $\bar{X} = 0.255$. Test at the 5% level of significance:H_0: $\mu = 0.25$ versus H_1: $\mu > 0.25$. As in Problem 2, assume that the diameters of rivets produced by the recently installed machine are distributed as $N(\mu, (0.0015)^2)$. What is the power of the test at $\mu = 0.251$?

9.4 TESTS CONCERNING THE MEAN OF A NORMAL POPULATION HAVING UNKNOWN VARIANCE

9.4.1 Case of a Left-Tail Test

Suppose that $X \sim N(\mu, \sigma^2)$ but the value of σ^2 is unknown and we wish to test the hypothesis

$$H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu < \mu_0$$

on the basis of a sample of size n taken from $N(\mu, \sigma^2)$.

In this case, we proceed as follows: Let $n, \bar{X}$, and S^2 be the sample size, sample average, and sample variance, respectively. For the probability α of the type I error, we choose the critical region in the $(\bar{X}, S)$-plane as the set of pairs of values $(\bar{X}, S)$ for which

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{n-1,1-\alpha} = -t_{n-1,\alpha} \tag{9.4.1}$$

where $-t_{n-1,1-\alpha} = t_{n-1,\alpha}$ is the value of t_{n-1}, the Student t-variable with $(n-1)$ degrees of freedom that satisfies

$$P(t_{n-1} > t_{n-1,\alpha}) = \alpha \tag{9.4.2}$$

Thus, for this critical region, we have

$$P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -t_{n-1,\alpha}|\mu = \mu_0\right) = \alpha \tag{9.4.3}$$

with α the desired level of significance. If a sample of size n from $N(\mu, \sigma^2)$ has values of $(\bar{X}, S)$ that satisfy equation (9.4.1), we say that $\bar{X}$ is *significantly smaller than* μ_0 at $100\alpha\%$ level of significance.

The probability of a type II error of the test above for H_0 against alternatives in H_1 is given by

$$\beta(\mu_1) = P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > -t_{n-1,\alpha}|\mu = \mu_1\right) \tag{9.4.4}$$

which may be rewritten as

$$\begin{aligned}\beta(\mu_1) &= P\left(\frac{(\bar{X} - \mu_1) - (\mu_0 - \mu_1)}{S/\sqrt{n}} > -t_{n-1,\alpha}|\mu = \mu_1\right)\\ &= P\left(\frac{\bar{X} - \mu_1}{S/\sqrt{n}} > \frac{\mu_0 - \mu_1}{S/\sqrt{n}} - t_{n-1,\alpha}|\mu = \mu_1\right)\\ &= P\left(t_{n-1} > \frac{\mu_0 - \mu_1}{S/\sqrt{n}} - t_{n-1,\alpha}\right)\end{aligned} \tag{9.4.5}$$

The reader should note the striking similarities between equations (9.4.5) and (9.3.9). The probability in equation (9.4.5) can be evaluated by using one of the statistical packages for selected values of $\delta = (\mu_0 - \mu_1)/(\sigma/\sqrt{n})$. The power of the test, as discussed earlier is given by $\gamma(\mu_1) = 1 - \beta(\mu_1)$. Interestingly, it turns out that $\gamma(\mu_1)$ depends on the parameter of noncentrality for this problem, given by $\delta = (\mu_0 - \mu_1)/(\sigma/\sqrt{n})$. The value of δ is known as soon as we set σ at an assumed value.

Example 9.4.1 (Testing a one-sided hypothesis about μ when σ is unknown) *Four determinations of copper in a certain solution yielded an average $\bar{X} = 8.30\%$ with $S = 0.03\%$. If μ is the mean of the population of such determinations, test, at the 5% level of significance, the hypothesis*

$$H_0\colon \mu = 8.32 \text{ versus } H_1\colon \mu < 8.32$$

Solution: This is a left-sided test, so, as in previous work, low values of $\bar{X}$ give evidence that H_1 is true. But in this case, σ is unknown, so that a test statistic is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Now, because of normality and the fact that $n = 4$, under $H_0\colon\ T \sim t_3$, so the critical region is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{3,0.95} = -t_{3,0.05} = -2.353$$

The observed value of the test statistic is

$$\frac{8.30 - 8.32}{0.03/\sqrt{4}} = -1.333$$

which is *not* less than -2.353, and hence, we do not reject the hypothesis H_0.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do not have sufficient evidence to conclude that the mean copper determination is less than 8.32.

9.4.2 Case of a Right-Tail Test

The reader may use the work in Section 9.4.1 to find that for a right-sided test of the hypothesis

$$H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu > \mu_0$$

on the basis of a sample from $N(\mu, \sigma^2)$, where σ^2 is unknown, the critical region for a given α consists of the set of pairs of values $(\bar{X}, S)$ for which

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{n-1,\alpha} \tag{9.4.6}$$

and that the probability β of type II error is

$$\beta(\mu_1) = P\left(t_{n-1} < \frac{\mu_0 - \mu_1}{S/\sqrt{n}} + t_{n-1,\alpha}\right) \tag{9.4.7}$$

This probability can be calculated using one of statistical packages, for selected values of δ, where $\delta = (\mu_0 - \mu_1)/(\sigma/\sqrt{n})$.

9.4.3 The Two-Tail Case

Now suppose that we want to test the hypothesis

$$H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu \neq \mu_0$$

on the basis of a sample from $N(\mu, \sigma^2)$, where σ^2 is unknown, In this case, we choose the critical region as the set of pairs of values $(\bar{X}, S)$ for which

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{n-1,\alpha/2} \tag{9.4.8}$$

Since

$$P\left(\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{n-1,\alpha/2} | \mu = \mu_0 \right) = \alpha \tag{9.4.9}$$

we then have the probability of a type I error for the test (equation (9.4.8)) is α. The probability of the type II error, β, for this test is given by

$$\beta(\mu_1) = P\left(\frac{\mu_0 - \mu_1}{S/\sqrt{n}} - t_{n-1,\alpha/2} < t_{n-1} < \frac{\mu_0 - \mu_1}{S/\sqrt{n}} + t_{n-1,\alpha/2} \right) \tag{9.4.10}$$

If $(\bar{X}, S)$ does (does not) satisfy the inequality in equation (9.4.8), we say that $\bar{X}$ does (does not) differ significantly from μ_0 at the α level of significance, which is equivalent to the statement that the $100(1-\alpha)\%$ confidence interval $(\bar{X} \pm t_{n-1,\alpha/2}(S/\sqrt{n}))$ for μ does not (does) contain the value μ_0 specified as the value of μ under H_0.

Example 9.4.2 (Testing a two-tail hypothesis about μ when σ is unknown) *In Example 9.4.1, suppose that we want to test the hypothesis*

$$H_0\colon \mu = 8.32 \text{ versus } H_1\colon \mu \neq 8.32$$

at the 5% level of significance.

Noting that $n = 4$ and that $t_{3,0.025} = 3.182$, we find by using equation (9.4.8) that the critical region is the set of values of $\bar{X}$ and S for which

$$\left| \frac{\bar{X} - 8.32}{S/\sqrt{4}} \right| > 3.182 \tag{9.4.11}$$

The observed value of $(\bar{X}, S)$ is $(8.30, 0.03)$, and hence, the left-hand side of equation (9.4.11) is observed to be

$$\left| \frac{8.30 - 8.32}{0.03/\sqrt{4}} \right| = 1.333$$

The value 1.333 is less than 3.182, so we do not reject the hypothesis H_0.

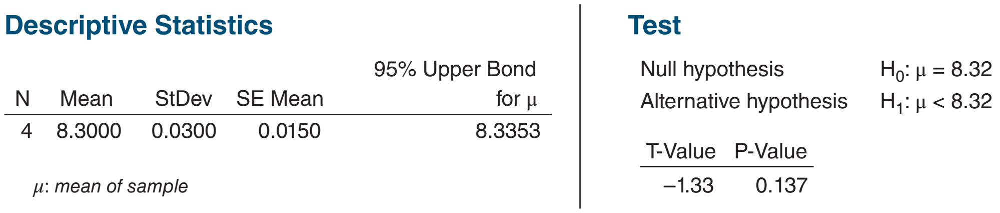
Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do not have sufficient evidence to conclude that the mean copper determination is different from 8.32.

The p-values for t-tests discussed in this section are again the areas under the left tail, right tail, or sum of the areas under both the tails of the t-distribution, depending on whether the test is left-tail, right-tail, or two-tail. Unlike the normal table, the t-table usually does not contain enough information to calculate the p-value accurately. The statistical packages discussed here are well equipped to evaluate the p-value accurately. We find, using MINITAB and R, the p-value and type II error for Example 9.4.1 in the following example, Example 9.4.3.

Example 9.4.3 (Using MINITAB and R) *Redo Example 9.4.1 using MINITAB and R. Find the probability of type II error* β *at* $\mu_1 = 8.31$.

MINITAB

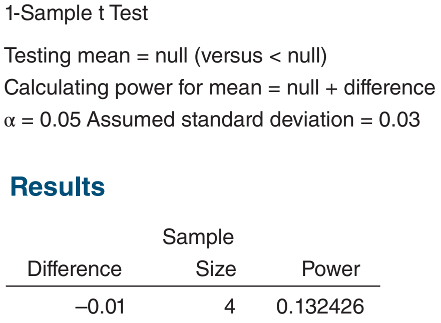
In Example 9.4.1, we have $\bar{X} = 8.30\%$ with $S = 0.03\%$ and $n = 4$. Follow the same steps as given in Example 9.3.4, except, in Step 3, we use Select **Stat** > **Basic Statistics** > **1-Sample t**. The MINITAB output that appears in the session window is:



```
Descriptive Statistics

                                    95% Upper Bond
N    Mean     StDev    SE Mean             for μ
4   8.3000   0.0300    0.0150             8.3353

     μ: mean of sample

Test

Null hypothesis          H0: μ = 8.32
Alternative hypothesis   H1: μ < 8.32

T-Value  P-Value
 –1.33    0.137
```

Since the p-value is 0.137, which is greater than the level of significance 0.05, we do not reject the null hypothesis.

We now proceed to find the power of the test and the probability of the type II error β at $\mu_1 = 8.31$ that is, we now find $\beta(\mu_1 = 8.31)$. Again using the same steps as given in Example 9.3.4, we obtain the MINITAB output shown here (note that $\mu_1 - \mu_0 = 8.31 - 8.32 = -0.01$.)



```
1-Sample t Test

Testing mean = null (versus < null)
Calculating power for mean = null + difference
α = 0.05 Assumed standard deviation = 0.03

Results

               Sample
  Difference     Size       Power
     –0.01          4    0.132426
```

Since the power is 0.132426, we then have that, β the probability of the type II error at $\mu_1 = 8.31$ is $\beta \approx 1 - 0.1324 = 0.8676$.

USING R

To conduct the one-sample t-test in R, the function 'tsum.test()' in R library 'BSDA' can be used. For the information provided in the Example 9.4.3, the t-test can be conducted by running the following in the R Console window.

```
install.packages(“BSDA”)
library(BSDA)
tsum.test(mean.x = 8.3, s.x = 0.03, n.x = 4, alternative = “less”, mu = 8.32)

#R output
One-sample t-Test
data: Summarized x
t = -1.3333, df = 3, p-value = 0.1373
alternative hypothesis: true mean is less than 8.32
```

Now, to find the probability of the type II error, first we calculate the power when $\mu = 8.31$. This can be done by using the ‘pwr.t.test()’ in ‘pwr’ library in R. Note that the effect size $d = (\mu_1 - \mu_0)/\sigma = (8.31 - 8.32)/0.03 = -1/3$ should be calculated to input to the R function. The following R code can be used to complete Example 9.4.3.

```
pwr.t.test(n = 4, d = -1/3, sig.level = 0.05, type = “one.sample”, alternative = “less”)

#R output
One-sample t test power calculation
n = 4
d = -0.3333333
sig.level = 0.05
power = 0.1324264
alternative = less
```

Since the power is 0.132426, we then have that the probability of the type II error at $\mu = 8.31$ is $\beta \approx 1 - 0.1324 = 0.8676$.

PRACTICE PROBLEMS FOR SECTION 9.4

1. Ten determinations of the percentage of water in a certain solution yielded $\bar{X} = 0.453\%$ and $S = 0.37\%$. If μ is the “true” percentage of water in the solution, assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 0.5$ versus H_1: $\mu < 0.5$.
2. A consumer group complains that the gas tax (in cents per gallon) levied by the federal, state, and local governments is too high. The following data give the gas tax (in cents per gallon) in 16 metropolitan areas around the country:

53	42	42	52	58	42	58	38
47	43	59	45	42	49	47	47

Assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 50$ versus H_1: $\mu > 50$.

3. The following data give the total cholesterol levels of 25 young male adults who are strict vegetarians:

99	97	110	110	117	91	90	120	113	120	103	115	104
90	95	98	91	93	101	94	112	108	103	120	119	

Assuming normality, test at the 1% level of significance the hypothesis H_0: $\mu = 110$ versus H_1: $\mu \neq 110$. Find the p-value.

4. The body temperature of infants rises significantly when they develop a cold. The following data give the body temperature of 16 infants measured 12 hours after the symptoms were first detected:

103	102	103	102	104	102	103	103
105	103	106	103	104	102	102	103

Assuming normality, test at the 1% level of significance the hypothesis H_0: $\mu = 103$ versus H_1: $\mu \neq 103$. Find the p-value.

5. The following data give the output voltages of a power supply:

13.76	13.97	13.94	13.81	14.92	13.77	12.64	13.52
13.27	13.83	12.68	14.33	12.81	12.63	12.46	13.98

Assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 13.5$ versus H_1: $\mu \neq 13.5$. Find the p-value.

6. The following data give the actual amount of beverage in sixteen “12-oz bottles”:

11.92	12.16	11.67	12.13	11.62	11.44	12.47	11.56
12.24	12.44	11.46	11.50	12.15	12.59	12.02	12.03

Assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 12$ versus H_1: $\mu \neq 12$. Find the p-value.

7. A method for determining the percent of impurity in various types of solutions is known to give determinations having a standard deviation of 0.03. Seven determinations on a certain chemical yield the values 7.18, 7.17, 7.12, 7.13, 7.14, 7.15, and 7.16. It is important that the chemical does not have more than 7.13% of impurities. Assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 7.13$ versus H_1: $\mu > 7.13$. State the critical region explicitly and state your conclusions.

8. A certain process yields pieces of steel wire with population standard deviation of their breaking strength equal to 500 psi. A random sample of nine test pieces of strands from the process yields $\bar{X} = 12,260$. If μ is the mean of the process, and assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 13,500$ versus H_1: $\mu \neq 13{,}500$. State your conclusions.

9.5 LARGE SAMPLE THEORY

In Sections 9.3 and 9.4, we studied tests of hypotheses on a single population mean under the assumption that the population under investigation is normal. In practice, however, situations may arise when the population under investigation is not normal. In the previous sections, we noted that tests of hypotheses on a single population mean are based on the sample mean $\bar{X}$. In Chapter 7, we also studied the famous theorem of statistical theory, the *central limit theorem*, which states that for sufficiently large sample sizes n ($n \geq 30$), the sample mean $\bar{X}$ is approximately normally distributed with mean μ and variance σ^2/n regardless of the population distribution, where, of course, μ and σ^2 are the population mean and variance. Thus, if the sample size is sufficiently large, to test hypotheses on a single population mean, we can use the results obtained in Section 9.3.

We summarize below in Table 9.5.1 the results applicable to test hypotheses on a single population mean when population is not normal, but the sample size is sufficiently large:

Table 9.5.1 Population variance known H_0: $\mu = \mu_0$.

Alternative hypothesis	Test statistic	Critical region
H_1: $\mu < \mu_0$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$
H_1: $\mu > \mu_0$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$
H_1: $\mu \neq \mu_0$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$\left\|\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right\| > z_{\alpha/2}$

Note that if the population standard deviation σ is unknown then the test statistics and the critical regions remain the same except that σ is replaced by the sample standard deviation S.

Example 9.5.1 (Testing the quality of tires) *A tire manufacturing company claims that its top-of-the-line tire lasts on average 65,000 miles. A consumer group tested 64 of these tires to check the claim. The data collected by this group yielded* $\bar{X} = 64{,}000$ *and standard deviation* $S = 4000$ *miles. Test at the* $\alpha = 0.05$ *level of significance the validity of the company's claim. Find the p-value of the test.*

Solution:

1. H_0: $\mu = 65{,}000$ versus H_1: $\mu < 65{,}000$.
2. $\alpha = 0.05$.
3. Since in this example, the population standard deviation is unknown, we use

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

as a suitable test statistic.

4. The sample size is 64, which is sufficiently large. Therefore, the test statistic

$$\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$$

is approximately distributed as standard normal if H_0 is true.
5. Since the test is left-tail with $\alpha = 0.05$, the rejection region is

$$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -z_{0.05} = -1.645$$

6. The value of the test statistic is

$$Z = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{64{,}000 - 65{,}000}{4000/\sqrt{64}} = -2.0$$

which falls in the rejection region. Thus, we reject the null hypothesis H_0 in favor of H_1. The p-value for the test is given by

$$p\text{-value} = P(Z < z) = P(Z < -2.0) = 0.0228$$

Since the p-value is smaller than $\alpha = 0.05$, we reject H_0.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do have sufficient evidence (p-value = 0.023) to conclude that on average the top-of-the-line tire lasts less 65,000 miles.

PRACTICE PROBLEMS FOR SECTION 9.5

1. A random sample of 50 water canteens of soldiers in a desert region show an average life of 4.34 years with a standard deviation of 1.93 years. Army experience is such that canteens are known to have an average life of 4.90 years. Test the hypothesis, at the 5% level of significance, that "canteen life" in the desert is truly 4.90 years, against the hypothesis that desert conditions decrease the life of the canteen.
2. Hourly wages of employees in a certain company have become a source of contention. An impartial arbitrator finds that industry wide hourly wages are approximately normally distributed with a mean of \$11.645 per hour. The arbitrator examines the earning records of 40 workers selected at random from the company's payroll list. He finds that the average is \$11.53, with a sample standard deviation of \$0.30. Test, at the 1% level of significance, the assertion (hypothesis) of the company that its wages conform to industry practices against the assertion (alternative hypothesis) that wages in this company are lower than that of the industry.
3. An aptitude test has been given over the past many years with a mean performance of 90. A group of 30 students are preparing for this test and are taught with special emphasis on remedial reading. The 30 students obtain an average of 94.2 with a sample standard deviation of 8.5. Has the remedial reading emphasis helped (at the 1% level of significance)?
4. A customer buys most of his wine from a winery in Napa Valley in California. He wants to check if the bottle-filling machine dispenses the exact amount of wine indicated on the bottle. He took a random sample of 36 bottles and measured the

actual amount of wine in them, each bottle containing, according to the label, 25.4 fluid ounces of wine. The data obtained are:

24.77	25.97	24.96	24.53	25.75	25.55	25.83	24.65	24.84	25.26	25.44	25.84
24.46	25.37	24.09	24.93	24.28	24.14	25.16	24.16	24.63	25.11	24.04	25.52
24.68	25.94	25.79	25.42	25.46	25.98	25.87	24.70	24.66	25.12	25.47	25.20

Assuming normality, test at the 5% level of significance the hypothesis $H_0\colon \mu = 25.4$ versus $H_1\colon \mu \neq 25.4$. Find the p-value.

5. Suppose in Problem 6 of Section 9.3 that the population standard deviation was not known. In order to test the hypothesis under consideration, the quality control engineer decided to take a random sample of 49 rivets and measure their diameters. The data obtained are as follows:

0.251	0.256	0.249	0.252	0.255	0.251	0.242	0.241	0.252	0.247	0.243	0.251	0.252
0.251	0.240	0.251	0.258	0.247	0.245	0.256	0.252	0.254	0.259	0.253	0.243	0.240
0.250	0.242	0.251	0.246	0.258	0.241	0.246	0.249	0.252	0.253	0.240	0.252	0.255
0.245	0.241	0.244	0.259	0.252	0.246	0.247	0.255	0.257	0.248			

Test at the 5% level of significance $H_0\colon \mu = 0.25$ versus $H_1\colon \mu > 0.25$. Find the p-value.

9.6 TESTS CONCERNING THE DIFFERENCE OF MEANS OF TWO POPULATIONS HAVING DISTRIBUTIONS WITH KNOWN VARIANCES

9.6.1 The Left-Tail Test

Suppose that $\bar{X}_1$ is the average of a random sample of size n_1 from a normal population $N(\mu_1, \sigma_1^2)$, and $\bar{X}_2$ is the average of a random sample of size n_2 from a normal population $N(\mu_2, \sigma_2^2)$, where σ_1^2 and σ_2^2 are both known. We know from Theorem 7.3.8 that $\bar{X}_1 - \bar{X}_2$ has the normal distribution $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. However, if the populations are not normal but if the sample sizes are sufficiently large, then $\bar{X}_1 - \bar{X}_2$ to good approximation is still distributed as $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$. We wish to test the hypothesis:

$$H_0\colon \mu_1 - \mu_2 = \delta_0 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 = \delta_1 < 0$$

on the basis of $\bar{X}_1 - \bar{X}_2$. This is a left-sided test based on the test statistic $\bar{X}_1 - \bar{X}_2$.

The procedure for testing this hypothesis is quite similar to that discussed in Section 9.3. If the probability of a type I error is to be α, it is evident from the results of Section 9.3 that the critical region for $\bar{X}_1 - \bar{X}_2$ is the set of values $\bar{X}_1 - \bar{X}_2$ for which

$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < -z_\alpha \tag{9.6.1}$$

Now under H_0: $\delta_0 = 0$, so that

$$P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < -z_\alpha\right) = \alpha \tag{9.6.2}$$

That is, the probability of type I error is α. The probability β of a type II error is the probability of accepting $\mu_1 - \mu_2 = \delta_0 = 0$ when $\mu_1 - \mu_2 = \delta_1 < 0$, so that

$$\beta = P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > -z_\alpha | \mu_1 - \mu_2 = \delta_1\right) \tag{9.6.3}$$

or

$$\beta = P\left(\frac{(\bar{X}_1 - \bar{X}_2) - \delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > -z_\alpha - \frac{\delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} | \mu_1 - \mu_2 = \delta_1\right) \tag{9.6.4}$$

which can be written as

$$\beta = 1 - \Phi\left(-z_\alpha - \frac{\delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right) = \Phi\left(z_\alpha + \frac{\delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right) \tag{9.6.5}$$

We note that if $\delta_0 \neq 0$ then equation (9.6.5) becomes

$$\beta = 1 - \Phi\left(\frac{\delta_0 - \delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} - z_\alpha\right) = \Phi\left(\frac{\delta_1 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} + z_\alpha\right) \tag{9.6.6}$$

which is the counterpart to β for the left-sided test in the one-population case in equation (9.3.9).

Returning to the case $\delta_0 = 0$, we have that if $\bar{X}_1 - \bar{X}_2$ satisfies equation (9.6.1), then we say that $\bar{X}_1 - \bar{X}_2$ is significantly less than zero or $\bar{X}_1$ is significantly smaller than $\bar{X}_2$ at the α level of significance, and/or that H_0 is rejected.

9.6.2 The Right-Tail Test

The case of a right-tail test can be treated in a similar manner as the left-tail test. The probability of the type II error in this case is given by

$$\beta = \Phi\left(\frac{\delta_0 - \delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} + z_\alpha\right) \tag{9.6.7}$$

We illustrate this case with the following example.

Example 9.6.1 (Testing two population means) *Two random samples have been obtained, one from each of population I and population II. The two populations have unknown means, but variances are known to be $\sigma_1^2 = 23.4$ and $\sigma_2^2 = 20.6$. The two samples yielded the following sample statistics:*

$$n_1 = 50, \bar{X}_1 = 38.5 \ \text{ and } \ n_2 = 45, \bar{X}_2 = 35.8$$

Test at the $\alpha = 0.05$ level of significance the hypothesis

$$H_0\colon \mu_1 - \mu_2 = \delta_0 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 = \delta_1 > 0$$

Solution: To test the above hypothesis, we proceed as follows:

1. H_0: $\mu_1 - \mu_2 = \delta_0 = 0$ versus H_1: $\mu_1 - \mu_2 = \delta_1 > 0$
2. $\alpha = 0.05$
3. The test statistic for the testing of the null hypothesis H_0: $\mu_1 - \mu_2 = \delta_0 = 0$ is
$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$
4. If the populations are not given to be normal but the sample sizes are sufficiently large (as in this example), thenby the central limit theorem and Theorem 7.3.8, we can easily show that the test statistic is approximately distributed as $N(0, 1)$. However, if the populations are given to be normal, then there is no restriction on the sample sizes.
5. Since the hypothesis in this example is right-tail, the rejection region is given by
$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > z_\alpha = 1.645$$
6. Substituting the values for $\bar{X}_1, \bar{X}_2, \sigma_1^2, \sigma_2^2$, and $\delta_0 = \mu_1 - \mu_2 = 0$ under the null hypothesis H_0: $\delta_0 = 0$, in the test statistic, the observed value of the test statistic is
$$\frac{(38.5 - 35.8) - 0}{\sqrt{23.4/50 + 20.6/45}} = 2.806.$$

This value is larger than 1.645, and hence, we reject the null hypothesis of equal means in favor of the alternative $\mu_1 - \mu_2 > 0$. In other words, based on the given information, we can conclude that at the $\alpha = 0.05$ level of significance, the mean of population I is greater than the mean of population II.

The p-value of the test using Table A.4 is equal to

$$p\text{-value} = P(Z \geq z) = P(Z \geq 2.806) = 0.0026$$

which is less than 0.05, the level of significance. Thus, we reject H_0.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do have sufficient evidence (p-value = 0.0026) to conclude that the mean of the first population is greater than that of the second population.

The probability of type II error β, say at $\delta_1 = \mu_1 - \mu_2 = 1$, as the reader may verify, is given by

$$\begin{aligned}\beta = \beta(1) &= \Phi\left(\frac{\delta_0 - \delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} + z_\alpha\right)\\ &= \Phi\left(\frac{-1}{\sqrt{23.4/50 + 20.6/45}} + 1.645\right)\\ &= \Phi(0.6057) = 0.7262\end{aligned}$$

Hence, the power of the test at $\delta_1 = 1$ is $\gamma(1) = 1 - \beta(1) = 1 - 0.7262 = 0.2738$.

9.6.3 The Two-Tail Test

In the two-sided case, the hypothesis to be tested is

$$H_0\colon \mu_1 - \mu_2 = \delta_0 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 = \delta_1 \neq 0$$

In this case, the critical region for the test at the α level of significance consists of the set of values of $\bar{X}_1 - \bar{X}_2$ for which

$$\left| \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right| > z_{\alpha/2} \tag{9.6.8}$$

The probability of type II error β is given by

$$\beta(\delta_1) = \Phi\left(\frac{\delta_0 - \delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} + z_{\alpha/2} \right) - \Phi\left(\frac{\delta_1 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} - z_{\alpha/2} \right) \tag{9.6.9}$$

The power of the test, that is, the probability of rejecting $\delta_0 = \mu_1 - \mu_2 = 0$ when $\delta_1 = \mu_1 - \mu_2 \neq 0$ is given by

$$\gamma(\delta_1) = 1 - \beta(\delta_1)$$

This may be written as

$$\gamma(\delta_1) = 1 - \left[\Phi\left(\frac{\delta_0 - \delta_1}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} + z_{\alpha/2} \right) - \Phi\left(\frac{\delta_1 - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} - z_{\alpha/2} \right) \right] \tag{9.6.10}$$

If $\bar{X}_1 - \bar{X}_2$ does (dose not) satisfy equation (9.6.8), we say that $\bar{X}_1$ does (does not) differ significantly from $\bar{X}_2$ at the α level of significance, which is equivalent to the statement that $\mu_1 - \mu_2 = 0$ is not (is) contained within the $100(1-\alpha)\%$ confidence interval $(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ for $\mu_1 - \mu_2$. We will discuss further the connection between the testing of hypotheses and the confidence intervals later in this chapter.

Example 9.6.2 (Testing two population means are equal) *Suppose two machines, say M_1 and M_2, are packaging 6-oz cans of talcum powder. It is known from the past behavior of the machines that the weights of their respective fillings are normal with standard deviations of 0.04 oz and 0.05 oz, respectively. Suppose 100 cans filled by each machine are emptied, the contents are carefully weighed, and the sample averages are $\bar{X}_1 = 6.11$ oz and $\bar{X}_2 = 6.14$ oz. We wish to test at the $\alpha = 0.01$ level of significance the hypothesis*

$$H_0\colon \mu_1 = \mu_2 \text{ versus } H_1\colon \mu_1 \neq \mu_2,$$

where μ_1 and μ_2 are means of populations of weights of fillings produced by machines M_1 and M_2, respectively.

Solution: We conduct the above testing of hypothesis as follows:

1. $H_0\colon \mu_1 = \mu_2$ versus $H_1\colon \mu_1 \neq \mu_2$.
2. $\alpha = 0.01$.

3. The test statistic for testing the hypothesis H_0: $\mu_1 = \mu_2$ or H_0: $\mu_1 - \mu_2 = 0$ is

$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

4. Using the same argument as in Example 9.6.1, the test statistic is distributed as $N(0,1)$, under H_0: $\delta_0 = 0$.
5. Since the hypothesis in this example is two-sided, the rejection region is given by

$$\left| \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right| > z_{\alpha/2}$$

6. Substituting the value of $\bar{X}_1, \bar{X}_2, \sigma_1^2, \sigma_2^2$, and the value of $\delta_0 = \mu_1 - \mu_2$, under the null hypothesis which is zero in the test statistic, the observed value of the test statistic is

$$\frac{6.11 - 6.14}{\sqrt{0.0016/100 + 0.0025/100}} = -4.685$$

which is smaller than -2.575 (very often 2.575 is used as an approximation of $Z_{0.005} = 2.576$), and hence, we reject the null hypothesis and accept the alternative hypothesis. The p-value of the test using the normal table is equal to

$$p\text{-value} = 2P(Z \geq |z|) = 2P(Z \geq 4.685) \approx 0$$

Since the p-value is less than 0.01, we reject the null hypothesis.

Interpretation: Based on the statistical testing procedure with significance level of 0.01, we do have sufficient evidence ($p - value \approx 0$) to conclude that the average filling weights of those two machines differ significantly.

Note that the 99% confidence interval for $\mu_1 - \mu_2$ is

$$(6.11 - 6.14) \pm 2.575\sqrt{0.0016/100 + 0.0025/100} = (-0.03 \pm 0.016) = (-0.046, -0.014),$$

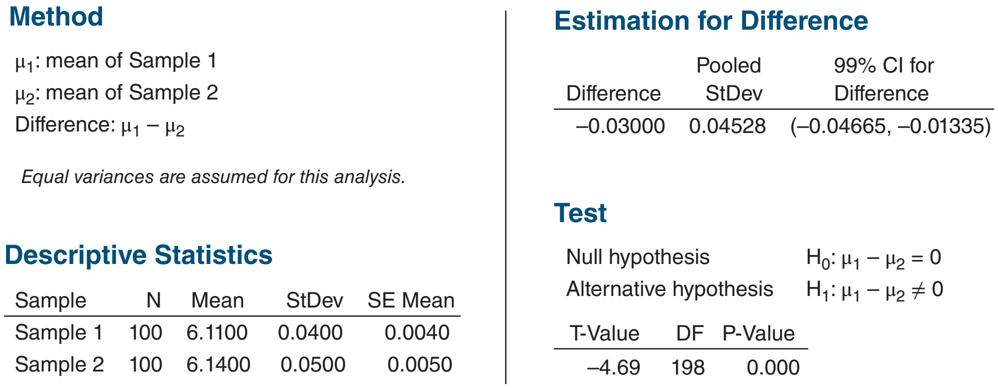
and does not contain the value 0.

Example 9.6.3 (Using MINITAB and R) *Do Example (equation (9.6.2)) using MINITAB and R.*

$$n_1 = 100,\ \bar{X}_1 = 6.11,\ \sigma_1 = 0.04$$
$$n_2 = 100,\ \bar{X}_2 = 6.14,\ \sigma_2 = 0.05$$

MINITAB

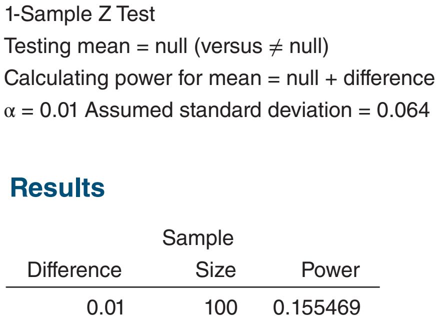
We follow the same steps as in Example (equation (9.3.4)), except, in Step 3, we use Select **Stat** > **Basic Statistics** > **2-Sample t**. Note that MINITAB does not have the option for a 2-Sample Z test. Thus, if we use **2-Sample t** when the populations are normal, variances are known, and the sample sizes are small, we will get only approximate results. However, in this example, even though the populations are normal and variances are known, we should get quite good results because the sample sizes are large. The MINITAB output that appears in the session window is:



```
Method                                        Estimation for Difference

μ1: mean of Sample 1                                              Pooled      99% CI for
μ2: mean of Sample 2                          Difference          StDev       Difference
Difference: μ1 − μ2                           -------------------------------------------------
                                                 −0.03000       0.04528   (−0.04665, −0.01335)
 Equal variances are assumed for this analysis.

                                              Test
Descriptive Statistics
                                              Null hypothesis          H0: μ1 − μ2 = 0
Sample       N     Mean     StDev   SE Mean   Alternative hypothesis   H1: μ1 − μ2 ≠ 0
------------------------------------------
Sample 1   100   6.1100   0.0400    0.0040     T-Value    DF   P-Value
Sample 2   100   6.1400   0.0500    0.0050    ---------------------------
                                                 −4.69   198     0.000
```

Since the p-value is 0.000, which is less than the level of significance 0.01, we reject the null hypothesis. That is, based on the present data, we can conclude that the packaging weights of cans of talcum powder by machines M_1 and M_2 are not the same.

We now proceed to find the power of the test and the type II error β at $\mu_1 - \mu_2 = 0.01$ (say), and for equal sample sizes, one can use $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ in Power and Sample Size for a 1-Sample Z procedure in MINITAB, so that here we use $\sigma = \sqrt{0.04^2 + 0.05^2} \approx 0.064$, and the MINITAB output is:



```
1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.01 Assumed standard deviation = 0.064

Results

                  Sample
Difference          Size      Power
-----------------------------------
      0.01           100   0.155469
```

P(type II error) $= \beta = 1 - 0.155469 = 0.844531$. For unequal sample sizes, one may use pooled standard deviation for power calculation in MINITAB, but result may differ from the true answer, and therefore, using the R approach is suggested.

USING R

Unlike in MINITAB, the exact two-sample z-test can be conducted in R. The built in R function ‘zsum.test()’ in library ‘BSDA’ can be used for this purpose. For the information provided in Example 9.6.3, the two-sample z-test can be conducted by running the following in the R Console window after installing the R library ‘BSDA’.

```
install.packages(“BSDA”)
library(BSDA)
zsum.test(mean.x = 6.11, sigma.x =. 04, n.x = 100, mean.y = 6.14, sigma.y = .05,
n.y = 100, alternative = “two.sided”, mu = 0, conf.level = 0.99)

#R output
Two-sample z-Test
data: Summarized x and y
z = -4.6852, p-value = 2.797e-06
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
-0.04649336 -0.01350664
sample estimates:
mean of x mean of y
6.11 6.14
```

Since the p-value is less than 0.01, we have ample evidence to reject the null hypothesis as we did using MINITAB.

To find the power, we use equation (9.6.10), and it can be implemented in R as follows:

```
power = 1-pnorm(-.01/sqrt(.04^2/100+.05^2/100) + abs(qnorm(0.01/2)))
         -pnorm(-.01/sqrt(.04^2/100+.05^2/100) - abs(qnorm(0.01/2)))

power
#R output
[1]0.155252
```

PRACTICE PROBLEMS FOR SECTION 9.6

1. A method for determining the percentage of iron in mixed fertilizer is available, and long experience with the method shows that its determinations are normally distributed with standard deviation of 0.12%. A company producing a certain type of fertilizer wishes to compare the findings of its laboratory with those of a state laboratory. The results are: $(n_1 = n_2 = 3)$

Company lab	8.84%	8.86%	9.16%
State lab	8.78%	8.96%	8.62%

   Test at the 5% level of significance the hypothesis that both laboratories do equivalent analysis, against the hypothesis that the state laboratory has a downward bias relative to the company laboratory (assume that $\sigma_1 = \sigma_2 = 0.12\%$).
2. It is known from past experience that two machines, A and B, used in producing a certain type of thread have standard deviations 0.04 and 0.03, respectively. The settings of the two machines are changed, and the concern is whether they were both

set alike. To check this, samples of 10 pieces of thread from machine A and 15 pieces of thread from machine B are taken at random, and it is found that $\bar{X}_A = 25.34$ and $\bar{X}_B = 25.42$. Test the hypothesis H_0: $\mu_A = \mu_B$ versus H_1: $\mu_A \neq \mu_B$. at the 5% level of significance. Graph the power function of the test.

3. Suppose that random samples of 25 are taken from two large lots of bulbs, A and B, and that $\bar{X}_A = 1610$ hours and $\bar{X}_B = 1455$ hours. Assuming that the standard deviation of bulb lives is 200 hours, test at the 5% level of significance the hypothesis H_0: $\mu_A - \mu_B = 120$ versus H_1: $\mu_A - \mu_B \neq 120$. Graph the power function. What is the power if $\mu_A - \mu_B = 100$?
4. Refer to Problem 8 of Section 8.4. The following data give the LDL cholesterol levels of two groups I and II of young female adults. Each member of group I follow a very strict exercise regimen, whereas in group II, no one does any exercise.

Group I	85	84	76	88	87	89	80	87	71	78
	74	80	89	79	87	75	83	84	71	70
Group II	91	105	98	98	98	107	101	101	94	96
	103	109	105	103	95	97	95	91	104	107

Suppose that μ_1 and μ_2 are means of two populations from which the young female adults in group I and group II have been selected. Assume that the two population are normally distributed with known variances of 35 and 30, respectively. Test at the 5% level of significance the hypothesis H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$. Find the p-value.

5. Refer to Problem 10 of Section 8.4. Two different brands of an all-purpose joint compound are used in residential construction, and their drying time, in hours, are recorded. Sixteen specimens for each were selected. Recorded drying times are as shown below.

Brand I	11.19	10.22	10.29	11.11	10.08	10.14	10.60	10.08
	11.28	11.98	11.22	11.97	10.47	10.79	11.98	10.03
Brand II	12.10	13.91	13.32	13.58	12.04	12.00	13.05	13.70
	12.84	13.85	13.40	12.48	13.39	13.61	12.37	12.08

Assume that the drying times of two brands are normally distributed with known variances of 0.5. Test at the 5% level of significance the hypothesis H_0: $\mu_I - \mu_{II} = 0$ versus H_1: $\mu_I - \mu_{II} \neq 0$. Find the p-value.

6. Two random samples from two normal populations, means μ_1 and μ_2, and with standard deviations $\sigma_1 = 4.5$ and $\sigma_2 = 6.2$, respectively, produced the following data:

Sample I	40	26	37	44	25	35	35	43	39	29	34	43	34	42	29	38
Sample II	51	40	29	47	43	36	47	38	40	26	26	38	37	27	34	35

(a) Test at the 2% level of significance the hypothesis H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$.
(b) Find the p-value for the test and make your conclusions using the p-value. Do you arrive at the same conclusion as in (a)?

9.7 TESTS CONCERNING THE DIFFERENCE OF MEANS OF TWO POPULATIONS HAVING NORMAL DISTRIBUTIONS WITH UNKNOWN VARIANCES

9.7.1 Two Population Variances are Equal

If the variances of σ_1^2 and σ_2^2 are unknown but are assumed to be equal, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ^2 is unknown, we proceed as follows:

The Left-Tail Test

The problem here is to test at the significance level α the hypothesis

$$H_0\colon \mu_1 - \mu_2 = \delta_0 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 = \delta_1 < 0$$

on the basis of the information in two samples, one from $N(\mu_1, \sigma^2)$ and one from $N(\mu_2, \sigma^2)$. We denote the sizes, averages, and variances of the two samples by $n_1, \bar{X}_1, S_1^2$ and $n_2, \bar{X}_2, S_2^2$, respectively. This is a left-tail test, and it is evident from Section 8.4 on the estimation of $\mu_1 - \mu_2$ that we can define the critical region for this test as the set of values $(\bar{X}_1, \bar{X}_2, S_1, S_2)$, for which

$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} < -t_{n_1+n_2-2,\alpha} \tag{9.7.1}$$

where under the null hypothesis $\delta_0 = 0$, and where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{9.7.2}$$

Note that when the sample sizes are equal, then

$$S_p^2 = \frac{S_1^2 + S_2^2}{2}$$

When H_0 is true, it is easily seen that the probability of a type I error when using equation (9.7.1) is α. The probability of a type II error β is given by

$$\beta = P\left(t_{n_1+n_2-2} > \frac{\delta_0 - \delta_1}{S_p\sqrt{1/n_1 + 1/n_2}} - t_{n_1+n_2-2,\alpha}\right) \tag{9.7.3}$$

with $\delta_0 = 0$.

The Right-Tail Test

We now consider a right-tail test for the hypothesis

$$H_0\colon \mu_1 - \mu_2 = \delta_0 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 = \delta_1 > 0$$

Clearly, the critical region for this test consists of the set of values of $(\bar{X}_1, \bar{X}_2, S_1, S_2)$ for which

$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} > t_{n_1+n_2-2,\alpha} \tag{9.7.4}$$

with $\delta_0 = 0$. The probability of making a type I error by this test is α. The probability of type II error β is given by

$$\beta = P\left(t_{n_1+n_2-2} < \frac{\delta_0 - \delta_1}{S_p\sqrt{1/n_1 + 1/n_2}} + t_{n_1+n_2-2,\alpha}\right) \tag{9.7.5}$$

with $\delta_0 = 0$.

The Two-Tail Test

If we wish to consider a two-tail test for the hypothesis

$$H_0\colon \mu_1 - \mu_2 = \delta_0 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 = \delta_1 \neq 0$$

then the critical region consists of the set of values of $(\bar{X}_1, \bar{X}_2, S_1, S_2)$ for which

$$\left|\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}}\right| > t_{n_1+n_2-2,\alpha/2} \tag{9.7.6}$$

with $\delta_0 = 0$. Again, we note that the probability of type I error is α. Here, β, the probability of type II error, is given by

$$\beta = P\left(\frac{\delta_0 - \delta_1}{S_p\sqrt{1/n_1 + 1/n_2}} - t_{n_1+n_2-2,\alpha/2} < t_{n_1+n_2-2} < \frac{\delta_0 - \delta_1}{S_p\sqrt{1/n_1 + 1/n_2}} + t_{n_1+n_2-2,\alpha/2}\right) \tag{9.7.7}$$

We again remind the reader that if $\sigma_1^2 = \sigma_2^2$ and if the total sample size $n_1 + n_2$ is large (in practice, $n_1 + n_2 \geq 62$), then $t_{n_1+n_2-2} \cong Z$. This means that the critical region, for example, defined by equation (9.7.6), may be stated for large $n_1 + n_2$ as

$$\left|\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}}\right| > z_{\alpha/2} \tag{9.7.8}$$

A similar remark applies to equations (9.7.1) and (9.7.4), and we may also modify the expressions defining β, the probability of the type II error accordingly by replacing $t_{n_1+n_2-2}$ by Z, the standard normal random variable.

Example 9.7.1 (Testing two population means when common variance is unknown) *Two methods of determining nickel content of steel, say M_1 and M_2, are tried on a certain kind of steel. Samples of four (4) determinations are made by each method, with the following results:*

$$\bar{X}_1 = 3.285\%,\ S_1^2 = 0.000033$$

$$\bar{X}_2 = 3.258\%,\ S_1^2 = 0.000092$$

As usual, we denote by μ_1 and μ_2 the means and by σ_1^2 and σ_2^2 the variances of the populations of determinations made by methods M_1 and M_2, respectively. We will assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and that nickel contents determined by the two methods are normally distributed. Suppose that we want to test, at the 5% level of significance, the hypothesis

$$H_0\colon \mu_1 = \mu_2 \text{ versus } H_1\colon \mu_1 > \mu_2$$

Solution: We conduct the above testing of hypothesis as follows:

1. $H_0\colon \mu_1 = \mu_2$ versus $H_1\colon \mu_1 > \mu_2$.
2. $\alpha = 0.05$.
3. The test statistic is
$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}}$$
4. Since the two populations are normal, the test statistic is distributed as Student t with $n_1 + n_2 - 2$ degrees of freedom, if $H_0\colon \mu_1 - \mu_2 = \delta_0 = 0$ is true.
5. This is a one-sided (right-tail) test with $n_1 + n_2 - 2 = 6$; hence the critical region is given by
$$\frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} > t_{6,0.05} = 1.943$$
6. The value of S_p^2 is
$$\frac{3(0.000033) + 3(0.000092)}{6} = 0.000063$$
which gives $S_p = 0.00794$, and here $\delta_0 = 0$. Therefore, the observed value of the test statistic is given by
$$\frac{0.027}{0.00794\sqrt{1/4 + 1/4}} = 4.809$$
which is greater than 1.943; hence, we reject the null hypothesis in favor of the alternative hypothesis that $\mu_1 > \mu_2$.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do have sufficient evidence to conclude that method 1 provides a higher average nickel content of steel than method 2.

Example 9.7.2 (Summary data on dimension of rotor shafts – using MINITAB and R) *Rotor shafts of the same diameter are being manufactured at two different facilities of a manufacturing company. A random sample of size $n_1 = 60$ rotor shafts from one facility produced a mean diameter of 0.536 in. with a standard deviation of 0.007 in. Another sample of size $n_2 = 60$ from the second facility produced a mean diameter of 0.540 in. with a standard deviation of 0.01 in. Assume that the two population variances are equal.*

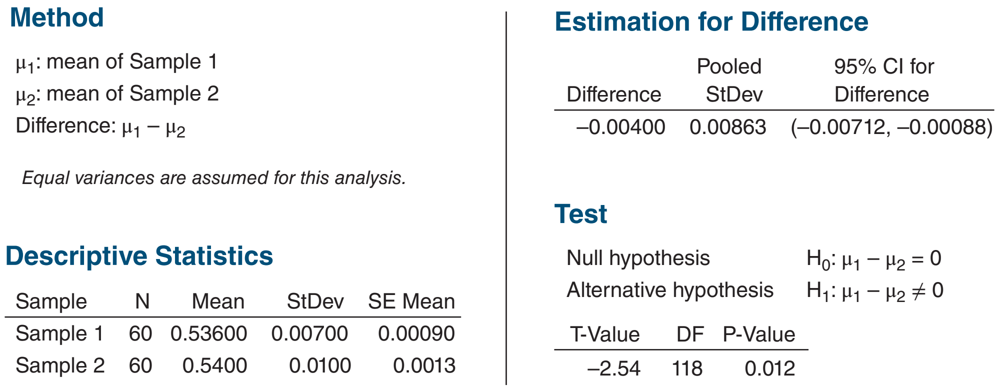
(a) Test the null hypothesis $H_0\colon \mu_1 - \mu_2 = 0$ versus $H_1\colon \mu_1 - \mu_2 \neq 0$ at the $\alpha = 0.05$ level of significance.
(b) Find the p-value for the test in (a).
(c) Find the size of the type II error β and the power of the test if the true value of $\mu_1 - \mu_2 = 0.002$.

MINITAB

In this example, we are given the following summary statistics:

$$n_1 = 60, \bar{X}_1 = 0.536, S_1 = 0.007 \text{ and } n_2 = 60, \bar{X}_2 = 0.540, S_1 = 0.010$$

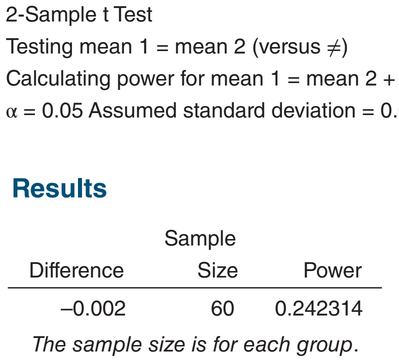
We again follow the same steps as in Example 9.3.4, except, in Step 3 we select **Stat** > **Basic Statistics** > **2-Sample t**, and select the option of equal variances. The MINITAB output that will appear in the session window is as shown below:



```
Method                                          Estimation for Difference

μ1: mean of Sample 1                                           Pooled       95% CI for
μ2: mean of Sample 2                            Difference      StDev       Difference
Difference: μ1 – μ2                             -----------------------------------------
                                                 –0.00400    0.00863   (–0.00712, –0.00088)
Equal variances are assumed for this analysis.

                                                Test
Descriptive Statistics
                                                Null hypothesis          H0: μ1 – μ2 = 0
Sample      N     Mean      StDev    SE Mean    Alternative hypothesis   H1: μ1 – μ2 ≠ 0
-----------------------------------------------
Sample 1   60   0.53600   0.00700   0.00090      T-Value    DF   P-Value
Sample 2   60    0.5400    0.0100    0.0013     --------------------------
                                                  –2.54    118     0.012
```

Since the p-value is 0.012, which is less than the level of significance 0.05, we reject the null hypothesis in favor of the alternative. That is, based on the given data, we conclude the two means differ significantly.

We now proceed to find the power of the test and the type II error probability, β if $\mu_1 - \mu_2 = -0.002$, and noting that the pooled standard deviation 0.0086. Again, using the same steps as given in Example 9.3.4, we get the MINITAB output as shown below:



```
2-Sample t Test
Testing mean 1 = mean 2 (versus ≠)
Calculating power for mean 1 = mean 2 + difference
α = 0.05 Assumed standard deviation = 0.00863


Results

                  Sample
  Difference       Size       Power
 -----------------------------------
     –0.002          60    0.242314
  The sample size is for each group.
```

Note that the MINITAB determines the power, only when both sample sizes are equal. For an estimate of σ, the standard deviation, use the value of pooled estimate, which in this example is 0.00863. The P(type II error) $= \beta = 1 - 0.242314 = 0.757686$.

USING R

The built in R function 'tsum.test()' in library 'BSDA' can be used to conduct two-sample t-test. For the information provided in Example 9.7.2, the test can be conducted by running the following in the R Console window.

```
install.packages(“BSDA”)
library(BSDA)
tsum.test(mean.x = 0.536, s.x = 0.007, n.x = 60, mean.y = 0.540, s.y = 0.01,
n.y = 60, alternative = “two.sided”, mu = 0, var.equal = TRUE, conf.level = 0.95)

#R output
Standard Two-Sample t-Test
data: Summarized x and y
t = -2.5383, df = 118, p-value = 0.01244
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.0071206309 -0.0008793691
sample estimates:
mean of x mean of y
0.536 0.540
```

Since the p-value is less than 0.05, we have sufficient evidence to reject the null hypothesis as we did using MINITAB.

To find the type II error, we first calculate the power when $\mu_1 - \mu_2 = -0.002$. This can be done by using the ‘pwr.t.test()’ function in R library ‘pwr’. Note that $d = (\mu_1 - \mu_2)/S_p = -0.002/0.00863 = -1/4.315$, where 0.00863 is the pooled standard deviation from the previous calculations. The following R code can be used to complete Example 9.7.2.

```
pwr.t.test(n = 60, d = -1/4.315, sig.level = 0.05, type = “two.sample”,
alternative = “two.sided”)

#R output
Two-sample t test power calculation
n = 60
d = 0.2317497
sig.level = 0.05
power = 0.2423142
alternative = two.sided
NOTE: n is number in *each* group
```

The P(type II error) $= \beta = 1 - 0.2423142 = 0.7576858$.

9.7.2 Two Population Variances are Unequal

If $\sigma_1^2 \neq \sigma_2^2$, we use the fact that for sufficiently large values of n_1 and n_2, the test statistic

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \tag{9.7.9}$$

is approximately distributed as $N(0, 1)$. Thus, by testing hypotheses, we can apply all the results obtained in Section 9.6 by replacing σ_1 and σ_2 with S_1 and S_2, respectively.

If n_1 and n_2 are small, equation (9.7.9) is approximately distributed as a Student t-variable with m degrees of freedom determined by Satterthwaite's approximation given in equation (8.4.12). Hence, for this case, we can apply all the results obtained so far in this section by changing the degrees of freedom from $n_1 + n_2 - 2$ to m. We illustrate this with the following example.

Example 9.7.3 (Testing equality of two population means when two population variances are unknown and unequal) *A new weight control company (A) claims that persons who use their program regularly for a certain period of time lose on average the same amount of weight as those who use the program of another well-established company (B) for the same period of time. Two random samples, the person who used company A's program and the second person who used company B's program, yielded the following summary statistics:*

$$n_1 = 12, \bar{X}_1 = 20, S_1 = 5, \quad \text{and} \quad n_2 = 10, \bar{X}_2 = 22, S_2 = 3$$

(The weight lost is measured in pounds). Test at the $\alpha = 0.01$ level of significance that the data do not provide sufficient evidence to support the claim of company A. Find the p-value of the test. We assume that the two populations are normally distributed with unequal variances.

Solution: We conduct the above testing of hypothesis as follows:

1. H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$.
2. $\alpha = 0.01$.
3. The test statistic for this problem is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

4. Here, the two populations are normal with unknown variances, and it is assumed that the variances are not equal. Thus in this case, the test statistic is distributed as Student's t with approximately m degrees of freedom, where

$$m = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} = \frac{8.9}{0.48457} = 18.37$$

so we use $m = 19$.
5. Since the test is a two-tail test and $\alpha = 0.01$, the rejection regions are given by

$$|T| > t_{19,0.005} = 2.861$$

6. Substituting the values $\bar{X}_1, \bar{X}_2, S_1, S_2$ and setting $\mu_1 - \mu_2$ under the null hypothesis H_0, we find that the observed value of the test statistic t is

$$T = \frac{(20 - 22) - (0)}{\sqrt{25/12 + 9/10}} \approx -1.16$$

which does not fall in the rejection region. Thus, we do not reject the null hypothesis H_0.

Since the test is a two-tail test, the p-value is given by p-value $= 2P(t \leq -1.16)$. From the t-table with 19 degrees of freedom, the reader can verify that

$$\begin{aligned} P(T \leq -1.328) &< P(T \leq 1.16) < \ P(T \leq -1.066) \\ 0.10 &< P(T \leq 1.16) < \ 0.15 \\ 2(0.10) &< 2P(T \leq 1.16) < 2(0.15) \\ 0.20 &< p\text{-value} < \quad 0.30 \end{aligned}$$

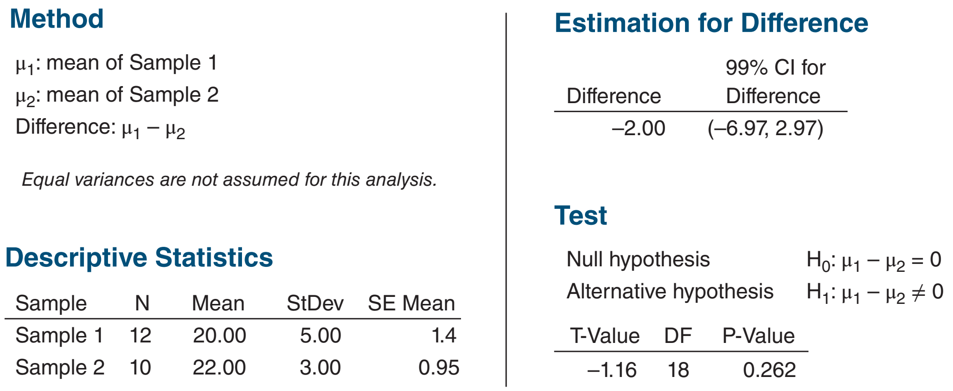
That is the p-value of the test is somewhere between 20% and 30%. Hence, we do not reject H_0. (The exact p-value can also be found using one of the statistical packages.)

Interpretation: Based on the statistical testing procedure with significance level of 0.01, we do not have sufficient evidence ($0.20 < p$-value < 0.30) to conclude that the mean weight loss of the two programs do differ.

We repeat the above example using MINITAB and R as follows:

MINITAB

We follow the same steps as in Example 9.3.4, except, in Step 3, we select **Stat** > **Basic Statistics** > **2-Sample t** and do not select the option of equal variances. The MINITAB output that will appear in the session window is as follows:



```
Method                                      Estimation for Difference

μ1: mean of Sample 1                                    99% CI for
μ2: mean of Sample 2                        Difference  Difference
Difference: μ1 − μ2                              −2.00  (−6.97, 2.97)

Equal variances are not assumed for this analysis.
                                            Test

Descriptive Statistics                      Null hypothesis         H0: μ1 − μ2 = 0
                                            Alternative hypothesis  H1: μ1 − μ2 ≠ 0
Sample     N   Mean   StDev  SE Mean
Sample 1  12  20.00    5.00      1.4        T-Value  DF  P-Value
Sample 2  10  22.00    3.00     0.95          −1.16  18    0.262
```

Since the p-value 0.262 is greater than the level of significance, which here is 0.01, we do not reject the null hypothesis in favor of the alternative. Previously, we had only a range for the p-value, whereas using MINITAB, we have the exact p-value. Furthermore, since the sample sizes are not equal, we cannot use MINITAB to find the power of the test and β, the probability of the type II error. However, the R can be used for this purpose.

USING R

The built in R function 'tsum.test()' in library 'BSDA' can be used to conduct two-sample t-test. For the information provided in the example 9.7.3, the test can be conducted by running the following in the R Console window after installing the R library ‘BSDA’. Make sure to use the option ‘var.equal = FALSE’ to get the Satterthwaite/Welch approximation.

```
tsum.test(mean.x = 20, s.x = 5, n.x = 12, mean.y = 22, s.y = 3, n.y = 10,
alternative = “two.sided”, mu = 0, var.equal = FALSE, conf.level = 0.95)

#R output
Welch Modified Two-Sample t-Test
data: Summarized x and y
t = -1.1579, df = 18.367, p-value = 0.2617
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.623586 1.623586
sample estimates:
mean of x mean of y
20 22
```

To find the type II error when $\mu_1 - \mu_2 = 1$, we can use the following manual R code. To complete Example 9.7.3, we use the given information.

```
#We assign numerical values for the pre-defined variables.
alpha = 0.05; df =18; diff = -1; sd1 = 5; n1 = 12; sd2 = 3; n2 = 10
ncp = (diff)/sqrt(sd1ˆ2/n1 + sd2ˆ2/n2)

#Probability of type II error
pt(qt(1-alpha/2,df),df, ncp)-pt(qt(alpha/2,df),df,ncp)

#R output
[1] 0.9148622
```

The P(type II error) $= \beta = 0.9148622$.

9.7.3 The Paired *t*-Test

In Sections 9.6 and 9.7, we studied the testing of hypotheses about the difference of two population means on the basis of two independent random samples, one sample from each population. Quite often, for various reasons, the experiments are designed in such a way that the data are collected in pairs, that is, two observations are taken on the same subject or experimental unit. Consequently, these two observations are not independent. We encounter these kind of data in various fields such as medicine, psychology, the chemical industry, and engineering. For example, a manager may want to evaluate the productivity of her workers before and after a training session; a nurse collects blood samples to test the serum-cholesterol level of patients before and after a treatment; a psychologist treats patients with similar mental disorders and takes two observations on each patient, one before and the other after treatment.

Data collected in this manner are usually known as *paired data*. Since the data collected have two observations on each subject, there is an implicit dependency between the two samples, one collected before an action and the other afterward. However, if we use the techniques of testing hypotheses discussed earlier in Section 9.7.1, where we had independent samples, then the results may turn out to be inaccurate.

Since the pairs of data are usually collected before and after a treatment, these data are sometimes called *before* and *after* data. The test method to test a hypothesis about the two means is called the *paired t-test.*

We now consider the problem that arises when we perform an experiment on n subjects, producing the n pairs of random variables $(Y_{11}, Y_{12}), (Y_{21}, Y_{22}), \ldots, (Y_{n1}, Y_{n2})$, where (Y_{i1}, Y_{i2}) is the pair of measurements obtained from the ith subject. It is assumed that the differences $Y_{i2} - Y_{i1} = X_i, \;\; i = 1, \ldots, n$, are n independent random variables all having identical normal distributions, namely $N(\mu, \sigma^2)$. Thus, the n differences $(X_1, \ldots, X_n)$ are essentially a random sample of size n from $N(\mu, \sigma^2)$. Let $\bar{X}$ and S^2 be the average and variance of this sample of X's. Then, the problem is to test the hypothesis $[\mu = E(X) = E(Y_1 - Y_2)]$.

$$H_0\colon \; \mu = \mu_0 \quad \text{versus} \quad H_1\colon \; \mu \neq \mu_0$$

The critical region for the test consists of the pairs $(\bar{X}, S)$ for which

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{n-1,\alpha/2} \tag{9.7.10}$$

which is a special case of the two-sided test discussed in Section 9.4, with $\mu_0 = 0$.

Example 9.7.4 (Testing equivalence of two methods using paired t-test) *Bennett and Franklin (1954) quote the following example of the use of the* t *test for a paired comparison problem. Table 9.7.1 gives the results of experiments on 21 different samples of iron ore using a standard dichromate titrimetric method and a new spectrophotometric method for the determination of the iron ore content of the 21 samples of iron ore. We wish to test* $\mu = 0$ *(at the 5% level of significance) against* $\mu \neq 0$, *where* μ *is the mean of the population of differences. From the data of Table 9.7.1, we find that* $\bar{X} = 0.0133$, $S = 0.0500$, *and hence, the test statistic in equation (9.7.10) has the value*

$$(0.0133)/(0.05/\sqrt{21}) = 1.22$$

Using Table A.5, we find $t_{20,.025} = 2.086$. Hence, we do not reject the null hypothesis that $\mu = 0$ and say that the present data show that the two methods do not differ significantly at the 5% level.

Before proceeding to the next section, we remind the reader that for sufficiently large degrees of freedom, say m, the Student t-variable with m degrees of freedom is approximately an $N(0,\,1)$ variable; that is for m large, $t_m \cong Z$. This means that if n is large, the critical regions of the tests discussed previously can be defined by replacing probability points of the t_{n-1} distribution with the corresponding point of the N(0, 1) distribution. For example, the critical region defined by equation (9.7.10) may be stated for sufficiently large n, as $|(\bar{X} - \mu_0)/(S/\sqrt{n})| > z_{\alpha/2}$, and so on.

Example 9.7.5 (Using MINITAB and R) *Use the data in Example 9.7.4 to test* $\mu = E(X_i) = E(Y_{i1} - Y_{i2}) = 0$ *(at the 5% level of significance) against* $\mu \neq 0$, *where* μ *is the mean of the population of differences. Find the* p*-value, the probability of type II error* β, *and the power of the test at* $\mu_1 = 0.03$.

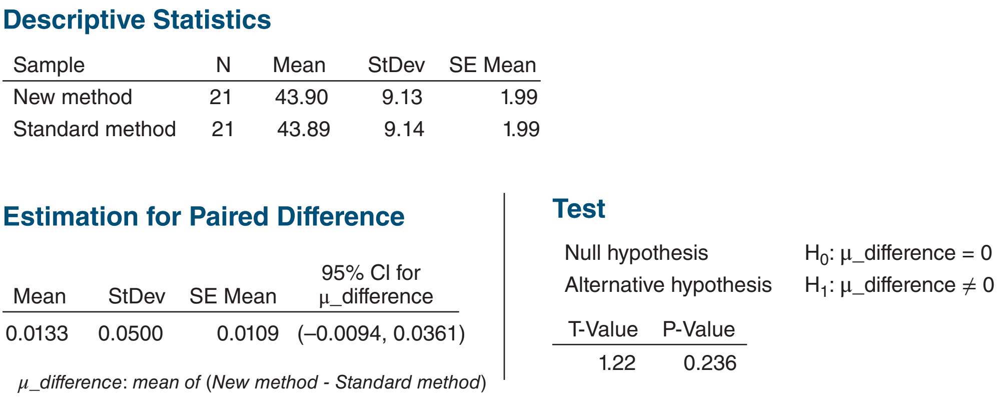
Table 9.7.1 Results of treating samples of iron ore with standard and new methods.

Sample	Standard method Y_{i1}	New method Y_{i2}	$X_i = Y_{i2} - Y_{i1}$
1	28.22	28.27	+0.05
2	33.95	33.99	+0.04
3	38.25	38.20	−0.05
4	42.52	42.42	−0.10
5	37.62	37.64	+0.02
6	36.84	36.85	+0.01
7	36.12	36.21	+0.09
8	35.11	35.20	+0.09
9	34.45	34.40	−0.05
10	52.83	52.86	+0.03
11	57.90	57.88	−0.02
12	51.52	51.52	0.00
13	49.49	49.52	+0.03
14	52.20	52.19	−0.01
15	54.04	53.99	−0.05
16	56.00	56.04	+0.04
17	57.62	57.65	+0.03
18	34.30	34.39	+0.09
19	41.73	41.78	+0.05
20	44.44	44.44	0.00
21	46.48	46.47	−0.01

MINITAB

To test a hypothesis for the mean μ at a given significance level, we proceed as follows:

1. Enter the data for the standard and new method in columns C1 and C2 of the Data window, respectively.
2. From the Menu bar, select **Stat** > **Basic Statistics** > **Paired t**. This will prompt a dialog box **Paired t for the Mean** to appear on the screen.
3. In the dialog box, select **Each sample is in a column** and select 'New method' as the **Sample 1** and 'Standard method' as the **Sample 2**.
4. Check **Options**, which prompts another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level**, enter zero next to **Hypothesized difference**, and enter the appropriate alternative hypothesis. In each dialog box, click **OK.** MINITAB output shows up in the Session window as:



Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
New method	21	43.90	9.13	1.99
Standard method	21	43.89	9.14	1.99

Estimation for Paired Difference

Mean	StDev	SE Mean	95% CI for μ_difference
0.0133	0.0500	0.0109	(–0.0094, 0.0361)

μ_difference: mean of (New method - Standard method)

Test

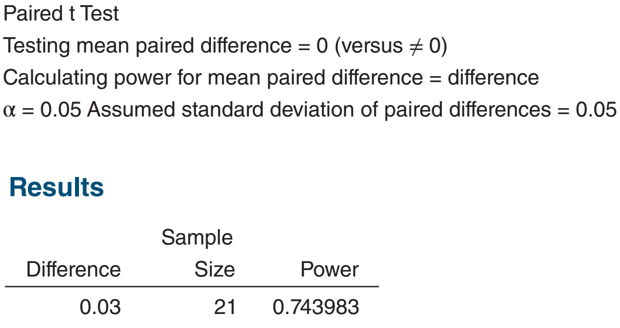
Null hypothesis H$_0$: μ_difference = 0
Alternative hypothesis H$_1$: μ_difference ≠ 0

T-Value	P-Value
1.22	0.236

Since the p-value is 0.236, which is greater than the level of significance 0.05, we do not reject H_0; that is the data do not indicate that the new method has shown any significant improvement or deterioration.

Now, we proceed to find the type II error β and the power of the test at $\mu_1 = 0.03$.

1. Select Stat > **Power and Sample Size** > **1-Sample t**, since this is a paired t-test. This will prompt a dialog box **Power and Sample Size for 1-Sample t** to appear on the screen.
2. In this dialog box, enter the value of **Sample size** (21), the **Difference** (0.03),and the value of **Standard deviation of paired differences** (0.05).
3. Select Options and make the necessary entries in the new dialog box that appears. In each dialog box, click **OK**. The Minitab output will show up in the Session window as given below:



Paired t Test
Testing mean paired difference = 0 (versus ≠ 0)
Calculating power for mean paired difference = difference
α = 0.05 Assumed standard deviation of paired differences = 0.05

Results

Difference	Sample Size	Power
0.03	21	0.743983

The type II error $\beta = 1$ - $0.743983 = 0.256017$.

USING R

The built in R function 't.test()' in library 'stats' can be used to conduct paired t-test. For the information provided in the example 9.7.5, the test can be conducted by running the following code in the R Console window after installing the 'stats' library in R.

```
Standard = c(28.22,33.95,38.25,42.52,37.62,36.84,36.12,35.11,34.45,
52.83,57.9,51.52,49.49,52.2, 54.04,56,57.62,34.3,41.73,44.44,46.48)
New = c(28.27,33.99,38.2,42.42,37.64,36.85,36.21,35.2,34.4,52.86,
57.88, 51.52,49.52,52.19,53.99, 56.04,57.65,34.39,41.78,44.44,46.47)

library(stats)
t.test(New, Standard, paired = TRUE)

#R output
Paired t-test
data: New and Standard
t = 1.2212, df = 20, p-value = 0.2362
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.009441562 0.036108229
sample estimates:
mean of the differences
0.01333333
```

Just as we found using MINITAB, the p-value is 0.236, which is greater than the level of significance 0.05, so that we do not have sufficient evidence to reject the null hypothesis that the means are equal.

To find the type II error, first we calculate the power when $\mu_1 = 0.03$. This can be done by using the ‘pwr.t.test()’ function in R library ‘pwr’. Note that $d = \mu_1/\text{std}(\text{New} - \text{Standard})$. The following R code can be used for this purpose.

```
pwr.t.test(n = 21, d = 0.03/sd(New-Standard), sig.level = 0.05,
type = “paired”, alternative = “two.sided”)

#R output
Paired t test power calculation
n = 21
d = 0.5996004
sig.level = 0.05
power = 0.7434236
alternative = two.sided
NOTE: n is number of *pairs*
```

The P(type II error) = β = 1-0.7434236 = 0.2565764. This answer is more accurate than the one we obtain from MINITAB as we did not round the standard deviation estimate.

PRACTICE PROBLEMS FOR SECTION 9.7

1. A machine is used to package “4-oz” boxes of a certain brand of gelatin powder. A modification is suggested to increase the speed of the operation, but there is some concern that the modified settings will cause the machine to fill the boxes with less powder than before. Accordingly, 50 boxes are filled before and after modification with the following results:

$$\begin{array}{ll}\text{Before}: & n_1 = 50,\ \bar{X}_1 = 4.091 \\ \text{After}: & n_2 = 50,\ \bar{X}_2 = 4.075\end{array}$$

Assuming that the machine yields packages whose weights are $N(\mu, (0.05)^2)$ for a wide range of values of μ, test at the 5% level of significance the hypothesis

$$H_0\colon\ \mu(\text{after}) = \mu(\text{before}) \quad \text{versus} \quad H_1\colon\ \mu(\text{after}) < \mu(\text{before})$$

Note that the two samples are independent. (*Note*: In this problem, variances are known.)

2. Nicotine determinations were made on each of six standard units of tobacco at each of two laboratories, say A and B, with the results shown below: (g = grams)

A : Nicotine content (g)	B : Nicotine content (g)
$26, 24, 28, 27, 32, 30$	$28, 31, 23, 29, 33, 32$

Test the hypothesis that the results of two laboratories are not significantly different at the 5% level of significance. Assume equal variances.

3. An experiment to determine the viscosity of two different brands of car oil, A and B, gives the results shown below. Test the hypothesis H_0: $\mu_A - \mu_B = 0$ at the 5% level of significance against the alternatives H_1: $\mu_A - \mu_B \neq 0$. Assume normality of the two populations with equal variances.

A: Viscosity	10.28	10.27	10.30	10.32	10.27	10.27	10.28	10.29
B: Viscosity	10.31	10.31	10.26	10.30	10.27	10.31	10.29	10.26

4. The following data give the productivity scores of 10 workers before and after a training program:

Before	95	97	105	94	103	97	98	95	100	95
After	105	111	111	106	106	104	105	102	106	102

Do these data provide sufficient evidence at the 5% level of significance that the training program has an effect? What assumption do you need to make in order to carry out this test?

5. The following data give the systolic blood pressures of 10 patients before and after their cardiologist switched them to a new drug therapy program:

Before	150	156	142	162	160	144	145	153	144	164
After	133	128	131	138	131	146	137	140	148	138

Do these data provide sufficient evidence at the 5% level of significance that the new drug therapy program is effective? What assumption do you need to make in order to carry out this test? Find the p-value.

6. Repeat Problem 2, assuming that the variances are not equal.
7. Repeat Problem 3, assuming that the variances are not equal.
8. Hemoglobin determinations were made on two sets of 11 animals that were exposed to two different environments. Each set had 11 animals. The data obtained is given below.

Set I	12.8	12.4	12.6	13.5	13.9	12.5	13.4	14.0	12.7	12.2	12.4
Set II	13.9	13.6	14.2	13.1	14.9	15.3	13.8	13.8	14.2	15.7	14.5

Test the hypothesis $H_0\colon \mu_1 - \mu_2 = 0$ at the 5% level of significance against the alternatives $H_1\colon \mu_1 \neq \mu_2$. Assume normality of the two populations, means μ_1 and μ_2, and with equal variances.

9.8 TESTING POPULATION PROPORTIONS

So far in this chapter, we have discussed methods of testing of hypotheses about population means. In this section, we discuss techniques of testing hypotheses about population proportions. In applications, it is quite common that we want to test such hypotheses. For example, we may be interested in verifying the percentage of the defective product manufactured by a company, or we may be interested in the percentage of the population of a country that is infected by HIV, or the proportion of employees of a company who are not happy with health insurance, or the proportion of students of a class who have made honors, or the proportion of drivers who are going above the posted speed limit on a given highway. We now proceed with testing a hypothesis about one population proportion; later in this section, we discuss methods of testing hypotheses about the difference of two population proportions.

9.8.1 Test Concerning One Population Proportion

Let $Y_1, Y_2, \dots, Y_n$ be a random sample from a dichotomous population or a population of Bernoulli trials with parameter p. Let $X = \Sigma Y_i$ be the total number of elements in the sample that possess the desired characteristic (success). From Chapter 8, we know that the sample proportion $\hat{p} = X/n$ is a point estimator of p. From Chapter 8, we also know that for large n $(np \geq 5,\ n(1-p) \geq 5)$ or if p unknown, $(n\hat{p} \geq 5,\ n(1-\hat{p}) \geq 5)$ the estimator $\hat{p}$ is distributed approximately by the normal distribution with mean p and variance $p(1-p)/n$. Using this result, we are now ready to discuss the testing of a hypothesis about the population proportion p. Under the assumption that the sample size is large, we discuss the following hypotheses about the population proportion:

$$H_0\colon\ p = p_0 \quad \text{versus} \quad H_1\colon\ p < p_0 \tag{9.8.1a}$$

$$H_0\colon\ p = p_0 \quad \text{versus} \quad H_1\colon\ p > p_0 \tag{9.8.1b}$$

$$H_0\colon\ p = p_0 \quad \text{versus} \quad H_1\colon\ p \neq p_0 \tag{9.8.1c}$$

Since the method of testing these hypotheses follows the same techniques that we used to test hypotheses about the population mean, we illustrate the method with the following example.

Example 9.8.1 (Testing a population proportion) *A set of environmentalists in the United States believe that sport utility vehicles (SUVs) consume excessive amount of gasoline and are thus "big" polluters of our environment. An environmental agency wants to find what proportion of vehicles on US highways are SUVs. Suppose that a random sample of 500 vehicles collected from highways in various parts of the country shows that 120 out of 500 vehicles are SUVs. Do these data provide sufficient evidence that 25% of the total vehicles driven in the US are SUVs? Use α= 0.05 level of significance. Find the p-value of the test.*

Solution: From the given information, we have

$$n = 500; X = \sum Y_i = 120$$

where Y_i is 1 if the vehicle spotted is an SUV and 0 otherwise, so that X is the total number of SUVs in the sample. Thus,

$$\hat{p} = X/n = 120/500 = 0.24$$

Now, to test the desired hypothesis, we take the following steps:

1. H_0: $p = p_0 = 0.25$ versus H_1: $p \neq 0.25$.
2. $\alpha = 0.05$.
3. We consider the pivotal quantity (see Chapter 8) for p as the test statistic, that is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \tag{9.8.2}$$

4. Now if H_0 is true, $np_0 = 500(0.25) = 125 > 5$, and $n(1 - p_0) = 500(1 - 0.25) = 375 > 5$, so that according to our criteria, the sample size is deemed large. The test statistic in equation (9.8.2) is therefore approximately distributed as standard normal $N(0, 1)$.
5. Since the test is a two-tail test and $\alpha = 0.05$, the rejection regions are given by

$$|Z| > z_{\alpha/2} = 1.96$$

6. Since $p_0 = 0.25$, and $\hat{p} = 0.24$, the value of the test statistic is

$$Z = \frac{0.24 - 0.25}{\sqrt{0.25(1 - 0.25)/500}} = -0.516,$$

which does not fall in the rejection region.
Thus, we do not reject the null hypothesis H_0.

Since the test is a two-sided, the p-value is given by

$$p\text{-value} = 2P(Z \geq |z|) = 2P(Z \geq 0.516) = 2(0.3030) = 0.6060$$

which is greater than the 5% level of significance, and we fail to reject the null hypothesis H_0. Thus, using the p-value, we arrive at the same decision.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do not have sufficient evidence (p-value = 0.606) to conclude that the percentage of SUVs on US highways is different from 25%.

Example 9.8.2 (Using MINITAB and R) *A civil engineer checked 50 concrete slabs for breaking strength. These concrete slabs were prepared by using a special mix. The engineer found that only 28 slabs either met or exceeded the desired standard of level of breaking strength. Does this summary statistic provide sufficient evidence at the 1% level of significance that less than 60% of the slabs would meet the desired standard? Find β, the probability of the type II error at $p_1 = 0.58$.*

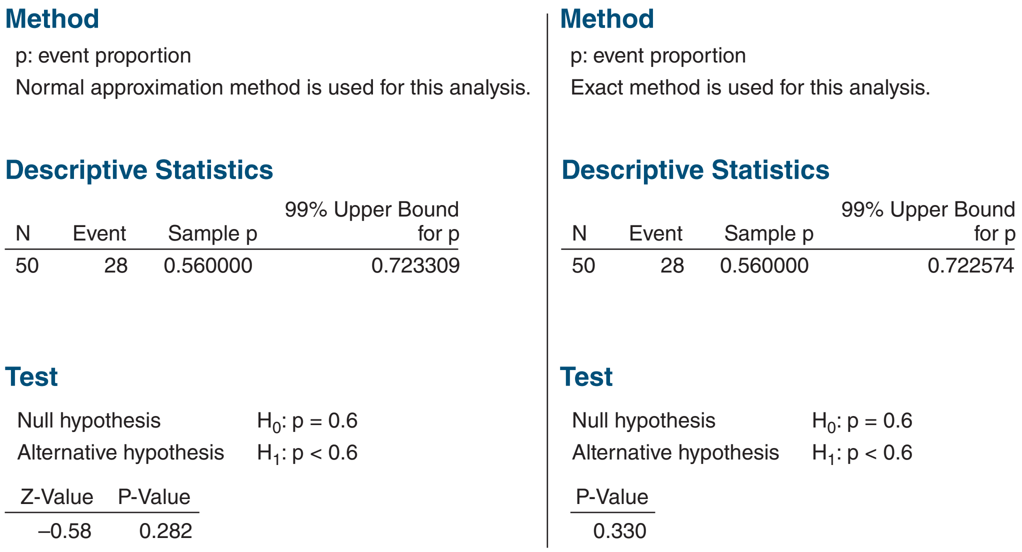
MINITAB

From the given information, we formulate the hypothesis as

$$H_0\colon\ p = 0.60 \quad \text{versus} \quad H_1\colon\ p < 0.60.$$

To test this hypothesis using MINITAB, we proceed as follows:

1. Select **Stat** > **Basic Statistics** > **1 Proportion.**This prompts a dialog box titled **One-Sample Proportion**to appear on the screen.
2. From the pulldown menu on the dialog box, select **Summarized data**. Then, enter **Number of events** (successes) and **Number of trials**.
3. Check the **Perform hypothesis test** and enter the value of the null hypothesis in the box next to **Hypothesized proportion**. Then, select **Options** and make the necessary entries in the new dialog box that appears. Make sure to select Exact or Normal approximation from the pulldown **Method** menu. In each dialog box, click **OK**. The Minitab output shows up in the Session window as given below.



```
Method
 p: event proportion
 Normal approximation method is used for this analysis.

Descriptive Statistics
                          99% Upper Bound
 N    Event   Sample p              for p
 50      28   0.560000           0.723309

Test
 Null hypothesis          H₀: p = 0.6
 Alternative hypothesis   H₁: p < 0.6

 Z-Value  P-Value
   –0.58    0.282
```

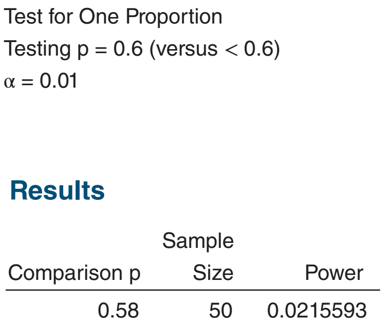
```
Method
 p: event proportion
 Exact method is used for this analysis.

Descriptive Statistics
                          99% Upper Bound
 N    Event   Sample p              for p
 50      28   0.560000           0.722574

Test
 Null hypothesis          H₀: p = 0.6
 Alternative hypothesis   H₁: p < 0.6

 P-Value
   0.330
```

Since the p-value is larger than the level of significance 1%, we do not reject H_0; that is the data provide insufficient evidence to conclude that the new mixture is producing less than 60% of the slabs that meet the standard.

To find the type II error β, we first find the power at $p_1 = 0.58$.

1. Select **Stat** > **Power and Sample Size** > **1 Proportion**. This prompts a dialog box **Power and Sample Size for 1 Proportion** to appear on the screen.

2. In this dialog box enter the value of **Sample sizes**, the **Comparison proportions**, and the **Hypothesized proportion**.
3. Select **Options** and make the necessary entries in the new dialog box that appears. Then, in each dialog box, click **OK.** The MINITAB output shows up in the Session window as given below:



```
Test for One Proportion
Testing p = 0.6 (versus < 0.6)
α = 0.01


Results

                      Sample
Comparison p     Size         Power
        0.58       50     0.0215593
```

Hence, the probability of the type II error is $\beta = 1 - power = 1 - 0.0215593 = 0.9784$.

USING R

The built in R functions 'prop.test()' and 'binom.test()' in library 'stats' can be used to conduct one sample normal approximation and exact binomial tests, respectively. For the information provided in Example 9.8.2, the test can be conducted by running the following R code in the Console window after installing the R library 'stats'.

```
#Normal approximation
prop.test(28, 50, p = .6, alternative = "less", conf.level = 0.99, correct = FALSE)
#Note: This test outputs the squared value of the test statistic,
i.e., X-squared = Z^2 in equation (9.8.2).

#R output
1-sample proportions test without continuity correction
data: 28 out of 50, null probability 0.6, X-squared = 0.33333,
df = 1, p-value = 0.2819, alternative hypothesis: true p is less than 0.6
99 percent confidence interval:
0.0000000 0.7093798
sample estimates:
p
0.56

#Exact Binomial test
binom.test(28, 50, p = 0.6, alternative = "less", conf.level = 0.99)

#R output
Exact binomial test
data: 28 and 50
number of successes = 28, number of trials = 50, p-value = 0.3299
alternative hypothesis: true probability of success is less than 0.6
```

```
99 percent confidence interval:
0.0000000 0.7225742
sample estimates:
probability of success
0.56
```

Just as we found using MINITAB, the p-values are 0.282 (normal approximation) and 0.330 (exact test), which are greater than the alpha-level of 0.01, so we do not have sufficient evidence to reject the null hypothesis.

To find the type II error, first we calculate the power when $p = 0.58$. This can be done by using the 'pwr.p.test()' function in R library 'pwr'. Note that the effect size h has to be calculated and inputted in the power calculation, which can be done by using the 'ES.h()' function. The following R code can be used to complete Example 9.8.2.

```
h = ES.h(0.58, .6)
pwr.p.test(h = h, n = 50, sig.level = 0.01, power = NULL, alternative = "less")

#R output
proportion power calculation for binomial distribution (arcsine transformation)
h = -0.04066727
n = 50
sig.level = 0.01
power = 0.02073565
alternative = less
```

The P(type II error) $=\beta$ = 1-0.02073565 = 0.97926435. This answer is slightly different from the one we obtain from MINITAB.

9.8.2 Test Concerning the Difference Between Two Population Proportions

Consider two binomial populations with parameters n_1, p_1 and n_2, p_2, respectively, where n_1 and n_2 are large. Then, we are usually interested in testing hypotheses such as

$$H_0\colon\ p_1 = p_2 \quad \text{versus} \quad H_1\colon\ p_1 < p_2 \tag{9.8.3a}$$

$$H_0\colon\ p_1 = p_2 \quad \text{versus} \quad H_1\colon\ p_1 > p_2 \tag{9.8.3b}$$

$$H_0\colon\ p_1 = p_2 \quad \text{versus} \quad H_1\colon\ p_1 \neq p_2 \tag{9.8.3c}$$

The above hypotheses can equivalently be written as

$$H_0\colon\ p_1 - p_2 = 0 \quad \text{versus} \quad H_1\colon\ p_1 - p_2 < 0 \tag{9.8.4a}$$

$$H_0\colon\ p_1 - p_2 = 0 \quad \text{versus} \quad H_1\colon\ p_1 - p_2 > 0 \tag{9.8.4b}$$

$$H_0\colon\ p_1 - p_2 = 0 \quad \text{versus} \quad H_1\colon\ p_1 - p_2 \neq 0 \tag{9.8.4c}$$

We illustrate the method of testing these hypotheses with the following example:

Example 9.8.3 (Testing equivalence of two population proportions) *A computer-assembly company gets all its chips from two suppliers. The company knows from experience in the past both suppliers have supplied a certain proportion of defective chips. The company wants to test alternative hypotheses: (a) supplier I supplies a smaller proportions of defective chips, (b) supplier I supplies a higher proportions of defective chips, or (c) the suppliers do not supply the same proportions of defective chips. To achieve this goal, the company took two random samples, one from the chips supplied by supplier I and the other from those supplied by supplier II. It was found that in the first sample of 500 chips, 12 were defective, and in the second sample of 600 chips, 20 were defective. Test each of the three hypotheses at the* $\alpha = 0.05$ *level of significance. Find the p-value for each test.*

Solution: From the given data, we have

$$n_1 = 500, \quad X_1 = 12, \quad \hat{p}_1 = X_1/n_1 = 12/500 = 0.024$$

$$n_2 = 600, \quad X_2 = 20, \quad \hat{p}_2 = X_2/n_2 = 20/600 = 0.033$$

where X_1 and X_2 are the number of defective chips in samples I and II, respectively.

To test the desired hypotheses, we proceed as follows:

1. (a) $H_0\colon p_1 - p_2 = 0$ versus $H_1\colon p_1 - p_2 < 0$
   (b) $H_0\colon p_1 - p_2 = 0$ versus $H_1\colon p_1 - p_2 > 0$
   (c) $H_0\colon p_1 - p_2 = 0$ versus $H_1\colon p_1 - p_2 \neq 0$
2. $\alpha = 0.05$.
3. We consider the pivotal quantity for $p_1 - p_2$ as the test statistic, that is

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \tag{9.8.5}$$

   with $p_1 - p_2 = 0$ under the null hypothesis.
4. Since $n_1\hat{p}_1 = 500(0.024) = 12 > 5$ and $n_1(1-\hat{p}_1) = 500(1-0.024) = 488 > 5$, we have reassurance in asserting that the sample size n_1 is large. Similarly, we can verify that the sample size n_2 can also be considered large. Thus, we assume that the test statistic in equation (9.8.5) is approximately distributed by the standard normal $N(0, 1)$.
5. Since the test statistic is approximately normally distributed, the rejection regions for testing hypotheses (a), (b), and (c) at the $\alpha = 0.05$ level of significance are given by

$$(a)\ Z < -z_\alpha \quad (b)\ Z > z_\alpha \quad (c)\ |Z| > z_{\alpha/2}$$

6. Since under the null hypothesis $p_1 - p_2 = 0$, we substitute the observed values of $\hat{p}_1, \hat{p}_2$ and $p_1 - p_2 = 0$ in the numerator of equation (9.8.5) and the values of p_1 and p_2 in the denominator. However, here p_1 and p_2 are unknown, but under the null hypothesis, $p_1 = p_2 = p$ (say), we can estimate p by pooling the two samples, that is

$$\hat{p} = (X_1 + X_2)/(n_1 + n_2) \tag{9.8.6}$$

We then replace p_1 and p_2 in the denominator with $\hat{p}$. In this example, we have

$$\hat{p} = (12 + 20)/(500 + 600) = 0.029$$

Thus, the value of the test statistic computed under the null hypothesis is given by

$$z = \frac{(0.024 - 0.0333) - 0}{\sqrt{0.029(0.971)/500 + (0.029)(0.971)/600}} = -0.92$$

Clearly, in all cases, the value of the test statistic does not fall in the rejection region. Thus, in either of the cases (a), (b), or (c), we do not reject the null hypothesis at the $\alpha = 0.05$ level of significance. In other words, the data imply, at the $\alpha = 0.05$ level of significance, that both the suppliers supply the same proportion of defective chips.

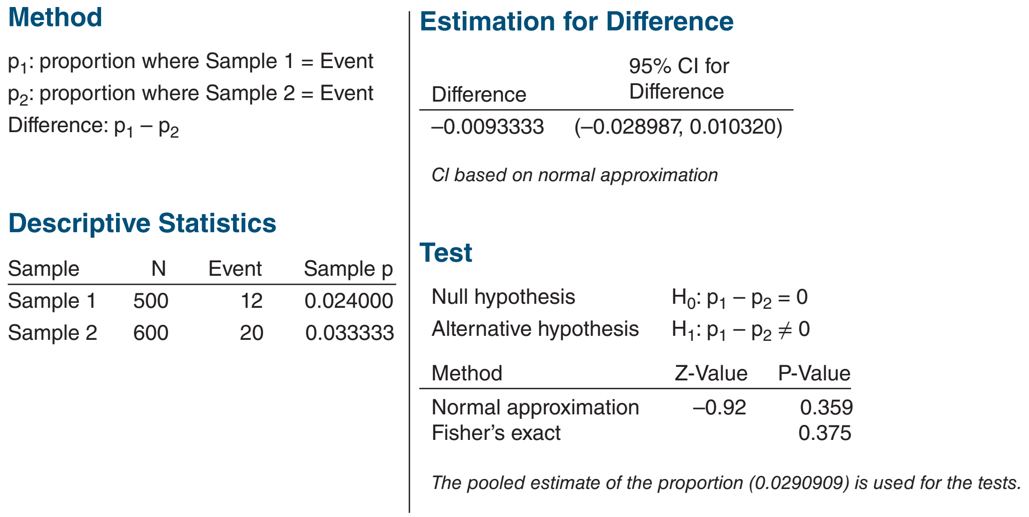
As a final remark, it is quite interesting to note that whether we test the hypothesis (a) or (b) or (c), all the steps including the value of the test statistic are exactly the same except for the rejection regions. However, the p-value of these tests will be different for different hypotheses. The p-value for each hypothesis is given by

(a) p-value $= P(Z \leq -0.92) \approx 0.1788$
(b) p-value $= P(Z \geq -0.92) \approx 0.82120$
(c) p-value $= 2P(Z \geq |-0.92|) \approx 2(0.1788) = 0.3576$

Example 9.8.4 (Using MINITAB and R) *Do Example 9.8.3 for case (c)-two-sided test-using MINITAB and R.*

MINITAB

1. Select **Stat** > **Basic Statistics** > **2 Proportions.** This prompts a dialog box **Two-Sample Proportion** to appear on the screen.
2. Select **Summarized data** from the pulldown menu. Enter number of events (12 and 20) and number of Trials (500 and 600) for Sample 1 and Sample 2, accordingly.
3. Select **Options** and make the necessary entries in the new dialog box that appears.
4. Select Use the pooled estimate of the proportion for **Test method**. Note that we check this option since we always test a hypothesis, by assuming the null hypothesis is true, which under the null hypothesis is that the two proportions are equal. Click **OK** on dialog boxes. The MINITAB output shows up in the session window as given below:



Method

p_1: proportion where Sample 1 = Event
p_2: proportion where Sample 2 = Event
Difference: $p_1 - p_2$

Descriptive Statistics

Sample	N	Event	Sample p
Sample 1	500	12	0.024000
Sample 2	600	20	0.033333

Estimation for Difference

Difference	95% CI for Difference
–0.0093333	(–0.028987, 0.010320)

CI based on normal approximation

Test

Null hypothesis	H_0: $p_1 - p_2 = 0$
Alternative hypothesis	H_1: $p_1 - p_2 \neq 0$

Method	Z-Value	P-Value
Normal approximation	–0.92	0.359
Fisher's exact		0.375

The pooled estimate of the proportion (0.0290909) is used for the tests.

Since the p-value is larger than the level of significance 5%, we do not reject H_0. Thus, at the α= 0.05 level of significance, there is no significant difference between the proportion of defective chips that both the suppliers supply.

USING R

The built in R function ‘prop.test()’ in library ‘stats’ can be used to conduct a two-sample proportion (normal approximation) test. For the information provided in Example 9.8.3, the test can be conducted by running the following code in the R Console window after installing the R library ‘stats’.

```
prop.test(x = c(12, 20), n = c(500, 600), alternative = c("two.sided"),
conf.level = 0.95, correct = FALSE)

#R output
2-sample test for equality of proportions without continuity correction
data: c(12, 20) out of c(500, 600)
X-squared = 0.84114, df = 1, p-value = 0.3591
alternative hypothesis: two.sided
95 percent confidence interval:
-0.02898696, 0.01032030
sample estimates:
prop 1 prop 2
0.02400000 0.03333333
```

Just as we found using MINITAB, the p-value is 0.359, and it is greater than the $\alpha = 0.05$ level of significance.

PRACTICE PROBLEMS FOR SECTION 9.8

1. Recall that for large n, the binomial distribution can be approximated by the normal distribution; that is for *large* n we may write, to good approximation,

$$Z \cong \frac{n\hat{p} - np}{\sqrt{np(1-p)}} \quad \text{or} \quad Z \cong \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

   where $\hat{p}$ is the observed sample proportion. What then, for n large, would be the approximate test of level α for testing (a) $p = p_0$ versus $p = p_1 > p_0$? (b) $p = p_0$ versus $p \neq p_0$?
2. A well-known politician claims he has 55% of the vote with him on a certain issue. A private poll of 1000 voters yields a sample proportion $\hat{p} = 0.51$. Could the well-known politician be right? State your level of significance.
3. During January, an electronics company produced 1350 printed circuits of which 146 were found to be defective. The following March, it produced 1300 circuits of which 113 were found to be defective. Assuming randomness, has the production process improved from January to March? Use $\alpha = 0.01$.
4. A dry-cleaning shop claims that a new spot remover will remove more than 70% of the spots to which it is applied. To check the claim, the spot remover is applied to 16 randomly selected spots. If only 12 of 16 spots are removed, test the following hypotheses: (a) $p = 0.70$ versus $p < 0.70$, (b) $p = 0.70$ versus $p \neq 0.70$. Use the level of significance $\alpha = 0.05$.
5. A heating oil company claims that one-fifth of the homes in a large city are heated by oil. Do we have reason to doubt this claim if, in a random sample 1000 homes in this city, it is found that 136 homes are heated by oil? Use a 0.01 level of significance.
6. An urban community wishes to show that the incidence of breast cancer is higher than in a nearby rural area. If it is found that 20 of 200 adult women in the urban community have breast cancer and 10 out of 150 adult women in the rural community have breast cancer, could we conclude at the 0.05 level of significance that breast cancer is more prevalent in the urban community?
7. A computer-assembly company gets all its chips from two suppliers. The company knows from experience that in the past both suppliers have supplied a certain proportions of defective chips. The company wants to test alternative hypotheses: (a) supplier I supplies smaller proportion of defective chips, (b) supplier I supplies higher proportions of defective chips, or (c) the suppliers do not supply the same proportion of defective chips. To achieve this goal, the company took two random samples, one from chips supplied by supplier I and the other from the ones supplied by supplier II. It was found that in the first sample of 500 chips, 20 were defective, and in the second sample of 600 chips, 30 were defective. For each of the above hypotheses, use $\alpha = 0.01$ as the level of significance. Find the p-value for each test.

9.9 TESTS CONCERNING THE VARIANCE OF A NORMAL POPULATION

Suppose that we have a random sample of size n from a normal distribution $N(\mu, \sigma^2)$ and that we wish to make a left-tail test of the hypothesis

$$H_0\colon \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1\colon \sigma^2 < \sigma_0^2$$

on the basis of the sample variance. Recalling from Theorem 7.3.5 that if $\sigma^2 = \sigma_0^2$, then $(n-1)S^2/\sigma_0^2$ is a chi-square random variable with $n{-}1$ degrees of freedom. We can use this fact as a test statistic. The critical region for the test is the set values of S^2 for which

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,\ 1-\alpha} \tag{9.9.1}$$

The probability of a type I error of this test is α.

In testing the hypothesis

$$H_0\colon\ \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1\colon\ \sigma^2 > \sigma_0^2$$

we have a right-tail test in which the critical region consists of the set of values of S^2 for which

$$\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,\ \alpha} \tag{9.9.2}$$

As is easily seen, the probability of type I error of this test is α.

In testing the hypothesis

$$H_0\colon \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1\colon \sigma^2 \neq \sigma_0^2$$

we have a two-tail test in which the critical region consists of the set of values of S^2 for which

$$\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1,\ 1-\alpha/2} \quad \text{or} \quad \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,\ \alpha/2} \tag{9.9.3}$$

The probability of a type I error for this test is again α.

Example 9.9.1 (Testing a population variance) *A sample of size 11 from a population that is normally distributed gives $S^2 = 154.6$. Test at the 5% level of significance the hypothesis*

$$H_0\colon\ \sigma^2 = 140 \quad \text{versus} \quad H_1\colon\ \sigma^2 > 140$$

This is a right-sided test. We note that $n = 11$ and $\alpha = 0.05$, so that we need $\chi^2_{10,.05} = 18.307$. Hence, from equation (9.9.2), the critical region is given by the values of S^2 for which

$$(10 \times S^2)/140 > 18.307 \tag{9.9.4}$$

But the observed value of $(10 \times S^2)/140 = 11.043$, which does not fall in the critical region (equation (9.9.4)), since 11.043 is not larger than 18.307. Therefore, based on the given data, we do not reject the null hypothesis H_0 that $\sigma^2 = 140$. The p-value is given by

$$p\text{-value} = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{10 \times 154.6}{140} | \sigma^2 = \sigma_0^2\right)$$

so that the p-value is

$$p\text{-value} = P(\chi^2_{10} > 11.043)$$

Using the chi-square table with 10 degrees of freedom, we can easily verify that

$$p\text{-value} = P(\chi^2_{10} > 11.043) > P(\chi^2_{10} > 15.9871) = 0.10 > 0.05$$

That is, as the p-value is greater than the significance level, so that we do not reject H_0.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do not have sufficient evidence (p-value > 0.10) to conclude that the population variance of the underlying distribution is higher than 140.

The power function of the tests described previously can be found in a straightforward manner. Consider the power of the test having critical region defined by equation (9.9.2).

The power is the probability of rejecting the hypothesis H_0 that $\sigma^2 = \sigma_0^2$ when $\sigma^2 = \sigma_1^2 > \sigma_0^2$ and is given by

$$\gamma(\sigma_1^2) = P\left(\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1,\alpha} | \sigma^2 = \sigma_1^2 > \sigma_0^2\right) = P\left(\frac{(n-1)S^2}{\sigma_1^2} > \frac{\sigma_0^2}{\sigma_1^2}\,\chi^2_{n-1,\alpha} | \sigma^2 = \sigma_1^2\right)$$

But when sampling from $N(\mu, \sigma_1^2)$, we have seen that $(n-1)S^2/\sigma_1^2$ is a χ^2_{n-1} variable. Hence, the power is

$$\gamma(\sigma_1^2) = P\left(\chi^2_{n-1} > \frac{\sigma_0^2}{\sigma_1^2}\,\chi^2_{n-1,\alpha}\right) \tag{9.9.5}$$

Thus, for instance, the power of the test in Example 9.9.1 at $\sigma_1^2 = 145$ is given by

$$\gamma(145) = P\left(\chi^2_{10} > \frac{140}{145} \times 18.307\right) = P(\chi^2_{10} > 17.6757) > 0.05$$

The probability of the type II error, β, at $\sigma^2 = \sigma_1^2$ is $1 - \gamma(\sigma_1^2)$. Thus, the probability of type II error $\beta(145)$ in Example 9.9.1 is less than 0.95.

Example 9.9.2 (Workers productivity data using MINITAB and R) *A production manager of a manufacturing company claims that over the last 10 years, worker productivity in the company did not vary more than 2%. To verify the manager's claim, the CEO of the company collected the information on worker's yearly productivity over the last 10-year period, as follows:*

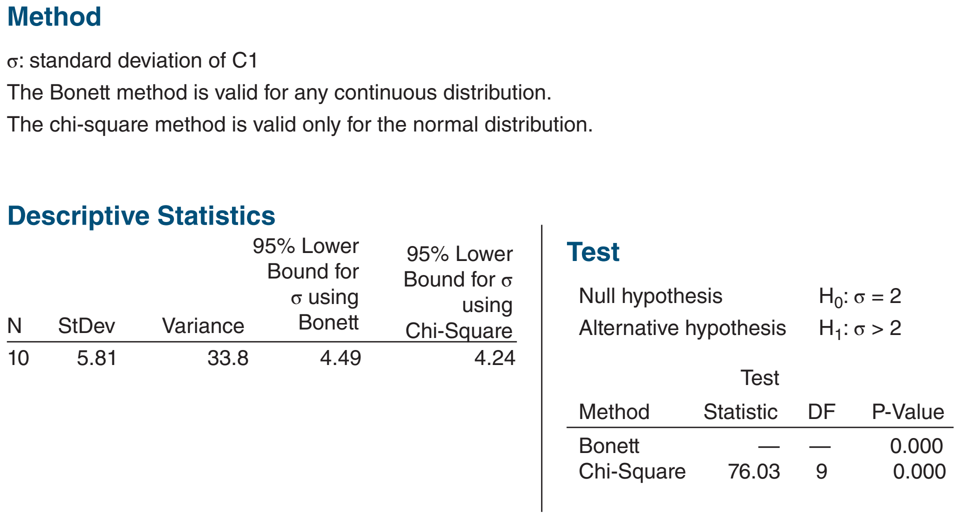
103 99 109 96 94 106 94 95 99 108

The manager wishes to test, at the 5% level of significance, the hypothesis

$$H_0\colon\ \sigma = \sigma_0 = 2 \quad \text{versus} \quad H_1\colon\ \sigma = \sigma_1 > 2$$

MINITAB

1. Enter the data in column C1 of the Worksheet.
2. Select **Stat** > **Basic Statistics** > **1 Variance.** This prompts a dialog box **One-Sample Variance** to appear on the screen.
3. From the pulldown menu, select **One or more samples, each in a column**, then make the appropriate entry in the box that follows.
4. Check the box next to **Perform hypothesis test** and enter the value of standard deviation (or variance) in the box next to **Hypothesized standard deviation (or variance)**. Then, select **Options** and make the necessary entries in the new dialog box that appears. The MINITAB output shows up in the session window given below:



```
Method

σ: standard deviation of C1
The Bonett method is valid for any continuous distribution.
The chi-square method is valid only for the normal distribution.


Descriptive Statistics                                    Test
                                                          
                          95% Lower     95% Lower         Null hypothesis         H₀: σ = 2
                          Bound for    Bound for σ        Alternative hypothesis  H₁: σ > 2
                           σ using        using
N   StDev   Variance       Bonett      Chi-Square                     Test
--------------------------------------------------        Method      Statistic  DF   P-Value
10   5.81       33.8         4.49            4.24         ------------------------------------
                                                          Bonett              —   —     0.000
                                                          Chi-Square      76.03   9     0.000
```

Note that the MINITAB delivers two results one using chi-square method and the other using (Bonett) an adjusted method. The chi-square method is for the normal distribution. The Bonett method is for any continuous distribution.

Since the p-value is less than the level of significance 5%, we reject the null hypothesis and conclude that the data do not support the manager's claim.

USING R

The built in R function 'sigma.test()' in library 'TeachingDemos' can be used to conduct chi-squared tests about population variances. Now for the information provided in Example 9.9.2, the test can be conducted by running the following code in the R Console window after installing the R library 'TeachingDemos'.

```
install.packages("TeachingDemos")
library(TeachingDemos)
x = c(103,99,109,96,94,106,94,95,99,108)
sigma.test(x, sigma = 2, alternative = "greater", conf.level = 0.95)
```

```
#R output
One sample Chi-squared test for variance
data: x
X-squared = 76.025, df = 9, p-value = 9.913e-13
alternative hypothesis: true variance is greater than 4
95 percent confidence interval:
17.9739 Inf
sample estimates:
var of x
33.78889
```

As in the MINITAB test, this p-value is well below the alpha-level, we have ample evidence to reject the null hypothesis.

PRACTICE PROBLEMS FOR SECTION 9.9

1. A producer claims that the diameters of pins he manufactures have a standard deviation of 0.05 inch. A sample of nine pins has a sample standard deviation of 0.07 inch. Is this sample value significantly larger than the claimed value of σ at the 5% level of significance?
2. Referring to the data of Problem 1 in Section 9.4, test at the 5% level of significance the hypothesis $\sigma = 0.4\%$ against the alternatives $\sigma < 0.4\%$.
3. Five determinations of percent of nickel in a prepared batch of ore produced the following results:

3.25	3.27	3.24	3.26	3.24

   Test, at the 5% level of significance, the hypothesis $\sigma = 0.01$ against the alternatives $\sigma > 0.01$.
4. Nine determinations were made by a technician of the melting point of manganese with the following results:

1268	1271	1259	1266	1257	1263	1272	1260	1256

   Test, at the 5% level of significance, the hypothesis $\sigma^2 = 40$ against the alternatives $\sigma^2 < 40$.
5. The following is the data of five independent replications of a chemical experiment:

7.27	7.24	7.21	7.28	7.23

   Test at the 5% level of significance the hypothesis $\sigma^2 = 0.9$ against the alternatives that $\sigma^2 \neq 0.9$.
6. The standard deviation S of muzzle velocities of a random sample of nine rounds of ammunition was found to be 93.2 ft/s. If the "standard" value of σ for the muzzle velocity of this type of ammunition is 70 ft/s, is the value of S significantly large at the 5% level of significance?

9.10 TESTS CONCERNING THE RATIO OF VARIANCES OF TWO NORMAL POPULATIONS

Suppose that we have two independent random samples of sizes n_1 and n_2 from two populations having the normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, and that we wish to test the hypothesis H_0 at significance level α:

$$H_0\colon\ \sigma_1^2/\sigma_2^2 = 1 \quad \text{versus} \quad H_1\colon\ \sigma_1^2/\sigma_2^2 = \lambda < 1$$

on the basis of independent samples of sizes n_1, n_2 and sample variances S_1^2, S_2^2. From Theorem 7.3.11, we know that $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ is a Snedecor F-ratio with $(n_1 - 1,\ n_2 - 1)$ degrees of freedom. Hence, the critical region for this left-sided test consists of the pairs of values of $(S_1^2,\ S_2^2)$ for which

$$S_1^2/S_2^2 < F_{n_1-1, n_2-1,\ 1-\alpha} \tag{9.10.1}$$

If $(S_1^2,\ S_2^2)$ satisfies equation (9.10.1), we say that S_1^2 is significantly smaller than S_2^2 at the α level of significance, and we reject the null hypothesis $H_0\colon \sigma_1^2/\sigma_2^2 = 1$ in favor of the alternative $H_1\colon \sigma_1^2/\sigma_2^2 = \lambda < 1$.

The critical region of the test for the hypothesis

$$H_0\colon\ \sigma_1^2/\sigma_2^2 = 1 \quad \text{versus} \quad H_1\colon\ \sigma_1^2/\sigma_2^2 = \lambda > 1$$

consists of the set of values of (S_1^2, S_2^2) for which

$$S_1^2/S_2^2 > F_{n_1-1, n_2-1,\ \alpha} \tag{9.10.2}$$

This is a right-sided test. If (S_1^2, S_2^2) satisfies equation (9.10.2), we say that S_1^2 is significantly larger than S_2^2 at the α level of significance, and we reject the null hypothesis $H_0\colon \sigma_1^2/\sigma_2^2 = 1$ in favor of the alternative $H_1\colon \sigma_1^2/\sigma_2^2 = \lambda > 1$.

The critical region of the two-sided test for the hypothesis

$$H_0\colon\ \sigma_1^2/\sigma_2^2 = 1 \quad \text{versus} \quad H_1\colon\ \sigma_1^2/\sigma_2^2 = \lambda \neq 1$$

consists of the set of values of (S_1^2, S_2^2) for which

$$S_1^2/S_2^2 < F_{n_1-1, n_2-1,\ 1-\alpha/2} \quad \text{or} \quad S_1^2/S_2^2 > F_{n_1-1, n_2-1,\ \alpha/2} \tag{9.10.3}$$

If (S_1^2, S_2^2) satisfies equation (9.10.3), we say that S_1^2 is significantly different from S_2^2 at the α level of significance, and we reject the null hypothesis $H_0\colon \sigma_1^2/\sigma_2^2 = 1$ in favor of the alternative $H_1\colon \sigma_1^2/\sigma_2^2 = \lambda \neq 1$.

The probability of making a type I error for each of the three tests discussed previously is α.

The power function of the left-tail, the right-tail, and two-tail tests previously described are seen to be, respectively,

$$\gamma(\lambda) = P\left(F_{n_1-1, n_2-1} < \frac{1}{\lambda} F_{n_1-1, n_2-1, 1-\alpha}\right), \quad \lambda < 1 \tag{9.10.4}$$

$$\gamma(\lambda) = P\left(F_{n_1-1,n_2-1} > \frac{1}{\lambda}F_{n_1-1,n_2-1,\alpha}\right), \quad \lambda > 1 \tag{9.10.5}$$

$$\gamma(\lambda) = P\left(F_{n_1-1,n_2-1} < \frac{1}{\lambda}F_{n_1-1,n_2-1,1-\alpha/2}\right) + P\left(F_{n_1-1,n_2-1} > \frac{1}{\lambda}F_{n_1-1,n_2-1,\alpha/2}\right), \quad \lambda \neq 1 \tag{9.10.6}$$

where $\lambda = \sigma_1^2/\sigma_2^2$.

Example 9.10.1 (Testing the equality of two population variances) *Five pieces of material were subjected to treatment T_1, and six pieces of a similar material were subjected to a different treatment, say T_2. Measurements made after T_1 and T_2 are applied gave the following results for the sample variances: $S_1^2 = 0.00045$ and $S_2^2 = 0.00039$. Test at the 5% level the hypothesis*

$$H_0\colon\ \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_1\colon\ \sigma_1^2 > \sigma_2^2$$

With respect to $\lambda = \sigma_1^2/\sigma_2^2$, this is a right-sided test, where we note that $\alpha = 0.05$, $n_1 = 5$, and $n_2 = 6$. Hence, we make use of $F_{4,5,.05} = 5.199$. Consulting equation (9.10.2), we have as the critical region of this test, the set of values of (S_1^2, S_2^2) for which

$$S_1^2/S_2^2 > 5.199.$$

But the observed value of S_1^2/S_2^2 in this example is $0.00045/0.00039 = 1.154$, which is not greater than 5.199. Hence, we do not reject the hypothesis $H_0\colon \sigma_1^2 = \sigma_2^2$.

Interpretation: Based on the statistical testing procedure with significance level of 0.05, we do not have sufficient evidence to conclude that the variability due to treatment T_1 is higher than that of treatment T_2.

Example 9.10.2 (Using MINITAB and R) *The CEO of the company in Example 9.9.2 now decides to compare his company's (say company 1) productivity with that of another company (say company 2). The productivity of the companies over the same period for the workers yielded the sample observations:*

Productivity 1:	103	99	109	96	94	106	94	95	99	108
Productivity 2:	95	94	105	98	105	95	104	100	105	101

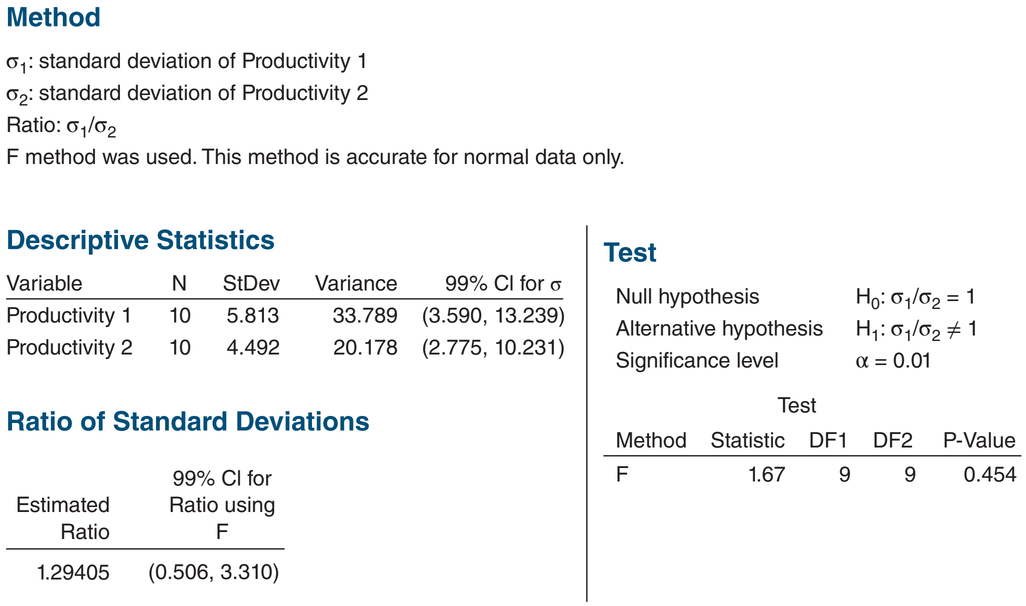
Test at the 1% level of significance the hypothesis

$$H_0\colon\ \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_1\colon\ \sigma_1^2 \neq \sigma_2^2$$

MINITAB

1. Enter the data in columns C1 and C2 of the Worksheet.
2. From the Menu bar, select **Stat** > **Basic Statistics** > **2-Variance.**This prompts a dialog box **Two-Sample Variance** to appear on the screen.

3. From the pulldown menu, select an appropriate option (for this example select **each sample is in its own column**) then make the appropriate entries in the boxes that follow.
4. Check **Options**, which prompts another dialog box to appear. Enter the desired confidence level in the box next to **Confidence level** and in the next to **Alternative hypothesis** select not equal. Make sure to check **Use test and confidence intervals based on normal distribution** otherwise it will output modified (robust) test results such as Bonett and Levene's. Click **OK** in each of the two dialog boxes. The MINITAB output shows up in the Session window as given below.



```
Method

σ1: standard deviation of Productivity 1
σ2: standard deviation of Productivity 2
Ratio: σ1/σ2
F method was used. This method is accurate for normal data only.

Descriptive Statistics

Variable        N   StDev   Variance    99% CI for σ
Productivity 1  10  5.813     33.789   (3.590, 13.239)
Productivity 2  10  4.492     20.178   (2.775, 10.231)

Ratio of Standard Deviations

                99% CI for
Estimated       Ratio using
    Ratio           F
1.29405      (0.506, 3.310)

Test

Null hypothesis          H0: σ1/σ2 = 1
Alternative hypothesis   H1: σ1/σ2 ≠ 1
Significance level       α = 0.01

                  Test
Method  Statistic  DF1  DF2  P-Value
F            1.67    9    9    0.454
```

Since the p-value is 0.454, which is greater than the level of significance 0.01, we do not reject the null hypothesis.

USING R

The built in R function 'var.test()' in library 'TeachingDemos' can be used to conduct an F-test for population variances. For the information provided in Example 9.10.2, the test can be conducted by running the following code in the R Console window after installing the R library 'TeachingDemos'.

```
x = c(103,99,109,96,94,106,94,95,99,108)
y = c(95,94,105,98,105,95,104,100,105,101)
var.test(x, y, ratio = 1, alternative = "two.sided", conf.level = 0.99)

#R output
F test to compare two variances
data: x and y
F = 1.6746, num df = 9, denom df = 9, p-value = 0.4543
```

```
alternative hypothesis: true ratio of variances is not equal to 1
99 percent confidence interval:
0.2560062 10.9534436
sample estimates:
ratio of variances
1.674559
```

As in the MINITAB test, this p-value is higher than the alpha-level, and we fail to reject the null hypothesis of equal variabilities.

PRACTICE PROBLEMS FOR SECTION 9.10

1. Resistance measurements were made on test pieces selected from two large lots L_1 and L_2, with the following results shown.

$L_1(\Omega)$	0.14	0.138	0.143	0.142	0.144	0.137
$L_2(\Omega)$	0.135	0.14	0.142	0.136	0.138	0.14

   If μ_1 and μ_2 are the means, and σ_1^2 and σ_2^2 are the variances of resistance measurements in L_1 and L_2, respectively, and assuming normality:
   (a) Test the hypothesis H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus H_1: $\sigma_1^2/\sigma_2^2 \neq 1$ at the 1% level of significance.
   (b) Using (a), test the hypothesis H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$ at the 5% level of significance.
2. Referring to the data of Problem 1, test at the 5% level of significance the null hypothesis H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus H_1: $\sigma_1^2/\sigma_2^2 < 1$.
3. Referring to Problem 8 of Section 8.4, test at the 5% level of significance the null hypothesis H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus H_1: $\sigma_1^2/\sigma_2^2 \neq 1$.
4. Referring to Problem 10 of Section 8.4, test at the 5% level of significance the hypothesis H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus H_1: $\sigma_1^2/\sigma_2^2 \neq 1$.
5. Two random samples from two normal populations with unknown standard deviations produced the following data:

Population 1	20	18	15	24	23	20	25	14	16	14
	16	24	22	16	20	15	22	16	20	20
Population 2	32	33	24	32	34	25	34	32	20	26
	29	21	22	37	27	30	24	22	22	30

   Test, at the 5% level of significance, the null hypothesis H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus H_1: $\sigma_1^2/\sigma_2^2 \neq 1$.
6. A manager in a manufacturing company wants to find if the workers' performance varies more in one shift than the other. She takes two random samples of 20 workers,

one sample each from each shift. Then, she observes the number of nonconforming parts produced during an eight-hour shift. The data obtained is given below:

Shift 1	10	8	9	8	9	10	12	11	9	10
	11	11	11	9	11	12	10	12	10	10
Shift 2	8	7	14	7	12	13	6	12	11	12
	8	7	6	7	8	13	6	14	8	12

Test, at the 5% level of significance, the null hypothesis H_0: $\sigma_1^2/\sigma_2^2 = 1$ versus H_1: $\sigma_1^2/\sigma_2^2 < 1$. What assumption must you make in order to carry out this test?

9.11 TESTING OF STATISTICAL HYPOTHESES USING CONFIDENCE INTERVALS

In Chapter 8, we studied certain techniques for constructing confidence intervals for population parameters, such as one population mean, difference of two population means, one population proportion, difference of two population proportions, one population variance, and the ratio of two population variances. Thus far, in this chapter, we have studied some techniques of testing hypotheses about these parameters. From our earlier discussion in this chapter and Chapter 8, the two techniques seen quite independent of each other, but the situation is in fact quite to the contrary: the two techniques are quite closely knitted together, in the sense that all the testing of hypotheses we have done so far in this chapter could have been done by using appropriate confidence intervals. We illustrate the concept of using confidence intervals by redoing some of the earlier examples in this chapter.

Example 9.11.1 (Testing a hypothesis using a confidence interval) *Recall that in Example 9.3.1, we considered a sample of 16 lengths of wire from a day's production of wire on a given machine, and the sample average was* $\bar{X} = 967.8$ psi. *The population of tensile strengths of wire in the day's production is* $N(\mu, \sigma^2)$*, and it is known from production experience that for this type of wire,* $\sigma = 128$ psi. *We wish to test the hypothesis*

$$H_0\colon\ \mu = \mu_0 = 1000 \quad \text{versus} \quad H_1\colon\ \mu = \mu_1 \neq 1000$$

Solution: Recall from Example 9.3.1 that the test statistic used for testing this hypothesis is $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$. It is clear that at the α significance level, we do not reject the null hypothesis H_0 if the value of the test statistic under the null hypothesis H_0: $\mu = \mu_0$ is such that

$$-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2} \tag{9.11.1}$$

or

$$-\frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \bar{X} - \mu_0 < \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

or

$$\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu_0 < \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \tag{9.11.2}$$

From equation (9.11.2), it follows that we do not reject the null hypothesis H_0 if the value μ_0 of μ under the null hypothesis falls in the interval $(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$.

This is equivalent to saying that we do not reject the null hypothesis H_0 if the confidence interval $(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$ for μ with confidence coefficient $1 - \alpha$ contains the value μ_0 of μ specified by the null hypothesis. Now, using the information contained in the sample summary, we obtain the confidence interval for μ with confidence coefficient $1 - \alpha = 0.95$ as

$$\begin{aligned}\left(\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\right) &= \left(967.8 - \tfrac{128}{\sqrt{16}}(1.96),\ 967.8 + \tfrac{128}{\sqrt{16}}(1.96)\right)\\ &= (905.08, 1030.52)\end{aligned}$$

Interpretation: With 95% confidence, the average tensile strength of that wire type is in between 905.08 and 1030.52 psi.

This interval clearly contains 1000, which is the value of μ under the null hypothesis. Thus, we do not reject the null hypothesis H_0, which is what we concluded in Example 9.3.2.

We now consider an example of a one-tail test.

Example 9.11.2 (Example 9.4.1 using a confidence interval) *Referring to Example 9.4.1, we have that four determinations of copper in a certain solution yielded an average* $\bar{X} = 8.30\%$ *with* $S = 0.03\%$. *If* μ *is the mean of the population of such determinations, we want to test, at the* $\alpha =$ *0.05 level of significance, the hypothesis*

$$H_0\colon\ \mu = \mu_0 = 8.32 \quad \text{versus} \quad H_1\colon\ \mu = \mu_1 < 8.32$$

using a one-sided confidence interval with 95% confidence coefficient.

Solution: Recall from Example 9.4.1 that the test statistic used to test this hypothesis is $(\bar{X} - \mu_0)/(S/\sqrt{n})$. Thus, it is clear that we do not reject the null hypothesis H_0 if the test statistic under the null hypothesis H_0: $\mu = \mu_0$ is such that

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > -t_{n-1,0.05} \tag{9.11.3}$$

or

$$\bar{X} - \mu_0 > -t_{n-1,0.05}\frac{S}{\sqrt{n}}$$

or

$$\mu_0 < \bar{X} + t_{n-1,0.05}\frac{S}{\sqrt{n}}$$

In other words, we do not reject the null hypothesis H_0 if the upper one-sided confidence interval

$$(-\infty, \bar{X} + t_{n-1,0.05}S/\sqrt{n}) \tag{9.11.4}$$

with confidence coefficient 95%, contains the value μ_0 of μ under the null hypothesis.

Now, using the information contained in the sample and the interval equation (9.11.4), we have that the upper one-sided confidence interval for μ with confidence coefficient 95% $(n = 4, \alpha = .05, t_{3,0.05} = 2.353)$ as

$$\begin{aligned}(-\infty, \bar{X} + t_{n-1,0.05}S/\sqrt{n}) &= (-\infty, 8.30 + 2.353(0.03)/\sqrt{4})\\ &= (-\infty, 8.335)\end{aligned}$$

Table 9.11.1 Confidence intervals with confidence coefficient $1-\alpha$ for testing various hypotheses.

Large sample sizes	
H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$	$(-\infty, \bar{X} + z_\alpha \sigma/\sqrt{n})$ if σ is known
H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$	$(-\infty, \bar{X} + z_\alpha S/\sqrt{n})$ if σ is unknown
H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$	$(\bar{X} - z_\alpha \sigma/\sqrt{n}, \infty)$ if σ is known
H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$	$(\bar{X} - z_\alpha S/\sqrt{n}, \infty)$ if σ is unknown
H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$	$(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$ if σ is known
H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$	$(\bar{X} - z_{\alpha/2}S/\sqrt{n}, \bar{X} + z_{\alpha/2}S/\sqrt{n})$ if σ is unknown
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 < 0$	$(-\infty, \bar{X}_1 - \bar{X}_2 + z_\alpha\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$ if σ_1, σ_2 are known
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 < 0$	$(-\infty, \bar{X}_1 - \bar{X}_2 + z_\alpha\sqrt{S_1^2/n_1 + S_2^2/n_2})$ if σ_1, σ_2 are unknown
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 > 0$	$(\bar{X}_1 - \bar{X}_2 - z_\alpha\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \infty)$ if σ_1, σ_2 are known
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 > 0$	$(\bar{X}_1 - \bar{X}_2 - z_\alpha\sqrt{S_1^2/n_1 + S_2^2/n_2}, \infty)$ if σ_1, σ_2 are unknown
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$	$(\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$ if σ_1, σ_2 are known
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$	$(\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{S_1^2/n_1 + S_2^2/n_2})$ if σ_1, σ_2 are unknown
Normal populations with small sample sizes	
H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$	$(-\infty, \bar{X} + z_\alpha \sigma/\sqrt{n})$ if σ is known
H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$	$(-\infty, \bar{X} + t_{n-1,\alpha} S/\sqrt{n})$ if σ is unknown
H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$	$(\bar{X} - z_\alpha \sigma/\sqrt{n}, \infty)$ if σ is known
H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$	$(\bar{X} - t_{n-1,\alpha} S/\sqrt{n}, \infty)$ if σ is unknown
H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$	$(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$ if σ is known
H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$	$(\bar{X} - t_{n-1,\alpha/2}S/\sqrt{n}, \bar{X} + t_{n-1,\alpha/2}S/\sqrt{n})$ if σ is unknown
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 < 0$	$(-\infty, \bar{X}_1 - \bar{X}_2 + z_\alpha\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$ if σ_1, σ_2 are known

(*continued overleaf*)

Table 9.11.1 (*continued*)

H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 < 0$	$(-\infty, \bar{X}_1 - \bar{X}_2 + t_{n_1+n_2-2,\alpha} S_p \sqrt{1/n_1 + 1/n_2})$, if σ_1, σ_2 are unknown and $\sigma_1 = \sigma_2$, where $S_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 < 0$	$(-\infty, \bar{X}_1 - \bar{X}_2 + t_{m,\alpha} \sqrt{S_1^2/n_1 + S_2^2/n_2})$ [for value of m see Section 9.7.2] if σ_1, σ_2 are unknown and $\sigma_1 \neq \sigma_2$
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 > 0$	$(\bar{X}_1 - \bar{X}_2 - z_\alpha \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \infty)$ if σ_1, σ_2 are known
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 > 0$	$(\bar{X}_1 - \bar{X}_2 - t_{n_1+n_2-2,\alpha} S_p \sqrt{1/n_1 + 1/n_2}, \infty)$ if σ_1, σ_2 are unknown and $\sigma_1 = \sigma_2$
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 > 0$	$(\bar{X}_1 - \bar{X}_2 - t_{m,\alpha} \sqrt{S_1^2/n_1 + S_2^2/n_2}, \infty)$ [for value of m see Section 9.7.2] if σ_1, σ_2 are unknown and $\sigma_1 \neq \sigma_2$
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$	$(\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2})$ if σ_1, σ_2 are known
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$	$(\bar{X}_1 - \bar{X}_2 \pm t_{n1+n2-2,\alpha/2} S_p \sqrt{1/n_1 + 1/n_2})$ if σ_1, σ_2 are unknown and $\sigma_1 = \sigma_2$
H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$	$(\bar{X}_1 - \bar{X}_2 \pm t_{m,\alpha/2} \sqrt{S_1^2/n_1 + S_2^2/n_2})$ if σ_1, σ_2 are unknown and $\sigma_1 \neq \sigma_2$
Binomial parameters; large sample sizes	
H_0: $p = p_0$ versus H_1: $p < p_0$	$(0, \hat{p} + z_\alpha \sqrt{\hat{p}(1-\hat{p})/n})$
H_0: $p = p_0$ versus H_1: $p > p_0$	$(\hat{p} - z_\alpha \sqrt{\hat{p}(1-\hat{p})/n}, 1)$
H_0: $p = p_0$ versus H_1: $p \neq p_0$	$(\hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n})$
H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 < 0$	$(-1, (\hat{p}_1 - \hat{p}_2) + z_\alpha \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)})$, where $\hat{p} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$
H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 > 0$	$((\hat{p}_1 - \hat{p}_2) - z_\alpha \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}, 1)$
H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 \neq 0$	$((\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)})$
	Note: Lower limit cannot be less than -1 and upper limit cannot be greater than 1.

(*continued overleaf*)

Table 9.11.1 (*continued*)

Normal populations but no restriction on sample sizes	
$H_0\colon \sigma^2 = \sigma_0^2$ versus $H_1\colon \sigma^2 < \sigma_0^2$	$\left(0, \dfrac{(n-1)S^2}{\chi^2_{(n-1),\ 1-\alpha}}\right)$
$H_0\colon \sigma^2 = \sigma_0^2$ versus $H_1\colon \sigma^2 > \sigma_0^2$	$\left(\dfrac{(n-1)S^2}{\chi^2_{(n-1),\ \alpha}}, \infty\right)$
$H_0\colon \sigma^2 = \sigma_0^2$ versus $H_1\colon \sigma^2 \neq \sigma_0^2$	$\left(\dfrac{(n-1)S^2}{\chi^2_{(n-1),\ \alpha/2}}, \dfrac{(n-1)S^2}{\chi^2_{(n-1),\ 1-\alpha/2}}\right)$
$H_0\colon \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ versus $H_1\colon \dfrac{\sigma_1^2}{\sigma_2^2} < 1$	$\left(0, F_{n_2-1,n_1-1,\ \alpha}\dfrac{S_1^2}{S_2^2}\right)$
$H_0\colon \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ versus $H_1\colon \dfrac{\sigma_1^2}{\sigma_2^2} > 1$	$\left(F_{n_2-1,n_1-1,\ 1-\alpha}\dfrac{S_1^2}{S_2^2}, \infty\right)$
$H_0\colon \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ versus $H_1\colon \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$	$\left(\dfrac{1}{F_{n_1-1,n_2-1,\ \alpha/2}}\dfrac{S_1^2}{S_2^2}, F_{n_2-1,n_1-1,\ \alpha/2}\dfrac{S_1^2}{S_2^2}\right)$

Interpretation: With 95% confidence, the average copper determination of that solution yields is at most 8.335%.

The confidence interval $(-\infty, 8.335)$ clearly contains the value 8.32, which is the value of μ under the null hypothesis. Thus, we do not reject the null hypothesis H_0, which is what we concluded in Example 9.4.1.

Having discussed these two examples, we now state the general rule and the confidence intervals to be used for testing various hypotheses discussed thus far in this chapter, see Table 9.11.1.

> ***The General Rule is***: *if the value of the parameter under the null hypothesis is contained in the corresponding confidence interval with confidence coefficient* $1-\alpha$, *then do not reject the null hypothesis* H_0 *at the* α *level of significance. Otherwise, reject the null hypothesis* H_0.

9.12 SEQUENTIAL TESTS OF HYPOTHESES

9.12.1 A One-Tail Sequential Testing Procedure

All tests considered in Sections 9.2–9.11 are based on samples of predetermined, fixed sample sizes n. In this section, we consider a procedure for testing the hypothesis

$$H_0\text{: population sampled has p.d.f. (or p.f.) } f_0(x) \quad \text{versus}$$
$$H_1\text{: population sampled has p.d.f. (or p.f.) } f_1(x)$$

where the sample size is not fixed in advance.

Such a procedure, called a *sequential test*, works as follows. We take a sequence of independent observations $X_1, X_2, \ldots,$ one at a time and make one of possibly three decisions after taking each observation. For the mth observation, $m = 1, 2, \ldots,$ the three possible decisions are as follows:

1. If
$$A < \frac{f_1(x_1)\cdots f_1(x_m)}{f_0(x_1)\cdots f_0(x_m)} < B \tag{9.12.1}$$
we draw an $(m+1)$st observation.
2. If
$$\frac{f_1(x_1)\cdots f_1(x_m)}{f_0(x_1)\cdots f_0(x_m)} \le A \tag{9.12.2}$$
we stop sampling and decide not to reject H_0.
3. If
$$\frac{f_1(x_1)\cdots f_1(x_m)}{f_0(x_1)\cdots f_0(x_m)} \ge B \tag{9.12.3}$$
we stop sampling and reject H_0 in favor of H_1.

The values of A and B are chosen so as to make the probability of type I and type II errors equal to α and β, respectively. Exact values of A and B are difficult to obtain. However, for the small values of α and β, ordinarily used in practice, we can use the well-known and fairly accurate approximations

$$A \approx \frac{\beta}{1-\alpha} \quad \text{and} \quad B \approx \frac{1-\beta}{\alpha} \tag{9.12.4}$$

Suppose that in Example 9.11.1 we have a normal population with mean μ and standard deviation $\sigma = 128$ and wish to test a hypothesis

$$H_0\colon\ \mu = \mu_0 \quad \text{versus} \quad H_1\colon\ \mu = \mu_1 > \mu_0$$

with $\alpha = 0.01$ and $\beta = 0.01$. The latter condition implies that the power of the test is to be 0.99; that is if the alternative hypothesis H_1 is true, we can reject H_0 with probability 0.99. Since the population is normal with mean μ and standard deviation σ, we have

$$f_0(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma}\right)^2\right], \; f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right]$$

On observing X_1, we find that

$$\frac{f_1(x_1)}{f_0(x_1)} = \exp\left[\frac{1}{2}\left(\frac{x_1-\mu_0}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma}\right)^2\right] = \exp\left[\left(\frac{\mu_1-\mu_0}{\sigma^2}\right)x_1 + \frac{\mu_0^2-\mu_1^2}{2\sigma^2}\right]$$

and if we observe $X_1, \ldots, X_m$, we find that

$$\frac{f_1(x_1)\times\cdots\times f_1(x_m)}{f_1(x_0)\times\cdots\times f_0(x_m)} = \exp\left[\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{m} x_i + \frac{m(\mu_0^2-\mu_1^2)}{2\sigma^2}\right]$$

Referring to equation (9.12.1) and equation (9.12.4) and the last expression above, we see that sampling continues as long as

$$\frac{\beta}{1-\alpha} < \exp\left[\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{m} x_i + \frac{m(\mu_0^2-\mu_1^2)}{2\sigma^2}\right] < \frac{1-\beta}{\alpha}$$

Taking natural logarithms, the above statement is equivalent to the following: sampling continues as long as

$$\ln\frac{\beta}{1-\alpha} < \frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{m} x_i + \frac{\mu_0^2-\mu_1^2}{2\sigma^2}m < \ln\frac{1-\beta}{\alpha}$$

where m is the number of observations already taken. Rearranging the terms in the last statement, we obtain that sampling continues as long as

$$\left(\frac{\sigma^2}{\mu_1-\mu_0}\right)\ln\frac{\beta}{1-\alpha} + \left(\frac{\mu_1+\mu_0}{2}\right)m < \sum_{i=1}^{m} x_i < \left(\frac{\sigma^2}{\mu_1-\mu_0}\right)\ln\frac{1-\beta}{\alpha} + \left(\frac{\mu_1+\mu_0}{2}\right)m \tag{9.12.5}$$

These inequalities may be displayed in a graph with $\sum_{i=1}^{m} x_i = T_m$ as ordinate and m as abscissa (see Figure 9.12.1). The lines having equations

$$T_m = \frac{\sigma^2}{\mu_1-\mu_0}\ln\frac{1-\beta}{\alpha} + \frac{\mu_1+\mu_0}{2}m \tag{9.12.6}$$

and

$$T_m = \frac{\sigma^2}{\mu_1-\mu_0}\ln\frac{\beta}{1-\alpha} + \frac{\mu_1+\mu_0}{2}m \tag{9.12.7}$$

are plotted, and as the observations are taken, T_m is plotted against m. Values for $\ln(1-\beta)/\alpha$ and $\ln\beta/(1-\alpha)$ for conventional values of α and β are displayed in Table 9.12.1.

Note that for this type of sequential sampling scheme, it is known that the expected sample size is usually appreciably less than the sample size required for testing H_0 versus H_1 with a sample of fixed size n that has the same α and β.

Table 9.12.1 Values of $\ln(1-\beta)/\alpha$ (upper entry in cell) and $\ln\beta/(1-\alpha)$, for various values of α and β.

α \ β	0.005	0.010	0.025	0.050	0.100
0.005	5.292	4.599	3.724	2.990	2.297
	−5.292	−5.288	−5.272	−5.247	−5.193
0.010	5.288	4.595	3.679	2.986	2.293
	−4.599	−4.595	−4.580	−4.554	−4.500
0.025	5.272	4.580	3.664	2.970	2.277
	−3.724	−3.679	−3.664	−3.638	−3.583
0.050	5.247	4.554	3.638	2.944	2.251
	−2.990	−2.986	−2.970	−2.944	−2.890
0.100	5.193	4.500	3.583	2.890	2.197
	−2.297	−2.293	−2.277	−2.251	−2.197

Example 9.12.1 (Sequential testing of a population mean) *Suppose that in Example 9.11.1 we want to test*

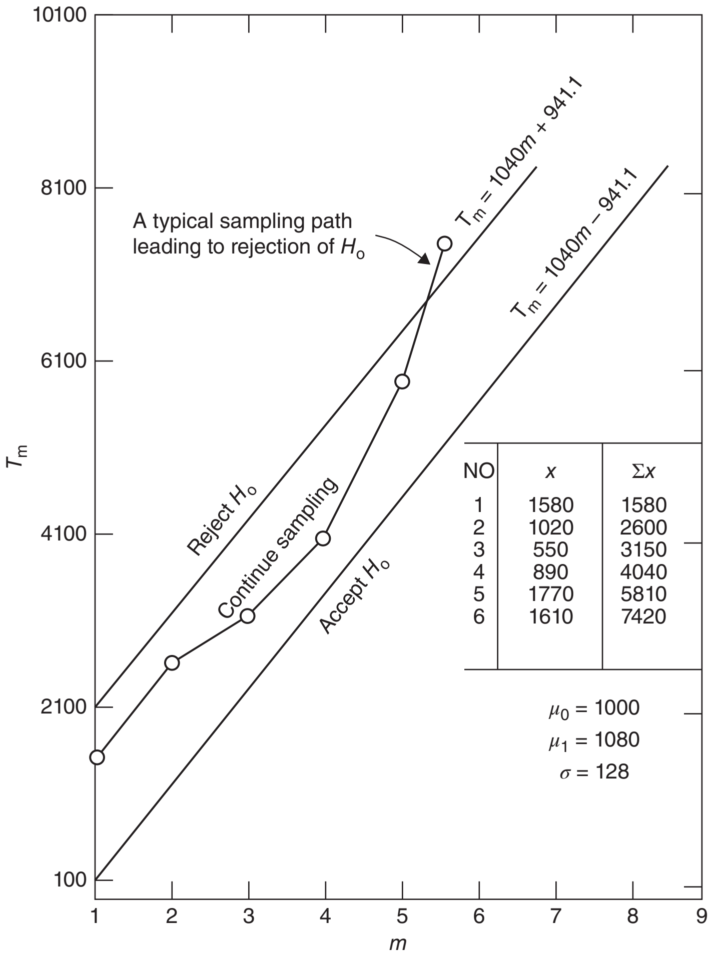
$$H_0\colon\ \mu = \mu_0 = 1000 \quad \text{versus} \quad H_1\colon\ \mu = \mu_1 = 1080$$

with $\sigma = 128$, $\alpha = 0.01$, *and* $\beta = 0.01$.

Then the two lines (equations (9.12.6) and (9.12.7)) have equations

$$T_m = 1040\,m + 941.1 \quad \text{and} \quad T_m = 1040\,m - 941.1$$

As long as the sequence of sample points $(m,\ T_m), m = 1, 2, \ldots,$ falls inside the band formed by these two lines, sampling is continued (see Figure 9.12.1). As soon as a point falls in the upper left region, the decision is made to stop sampling and reject H_0 in favor



NO	x	Σx
1	1580	1580
2	1020	2600
3	550	3150
4	890	4040
5	1770	5810
6	1610	7420

Figure 9.12.1 Decision regions for the sequential test of Example 9.12.1 showing a typical sampling path.

of H_1. As soon as a point falls in the lower right region, the decision is made to stop sampling, but not to reject H_0.

From Figure 9.12.1, it is clear that when $m = 6$, the value of T_m rises above the decision band. We are thus led to stop sampling and to reject the null hypothesis.

We can similarly define a sequential test if the random variable on which we make our sequence of independent observations is discrete, that is, if we use probability functions (p.f.s) $p_0(x)$ and $p_1(x)$ rather than probability density functions (p.d.f.s) $f_0(x)$ and $f_1(x)$.

Example 9.12.2 (Sequential testing of a population proportion) *A certain process yields mass-produced items that are* $100p\%$ *defective,* p *unknown. The manufacturer would like to conduct a sequential test of the hypothesis*

$$H_0\colon\ p = p_0 \quad \text{versus} \quad H_1\colon\ p = p_1 > p_0$$

at significance level α *and power* $1 - \beta$ *at* $p = p_1$*; that is, the desired value of the probability of type II error is to be* β *at* $p = p_1$*.*

Solution: Let X be a random variable having value 1 when an item drawn from the process is defective and value 0 when the item is nondefective. The probability function of X is given by

$$p(x) = p^x(1-p)^{1-x}, \qquad x = 0,\ 1$$

If we draw m items, and if the values of X obtained are $X_1, \ldots, X_m$, then

$$p(x_1)\cdots p(x_m) = p^{T_m}(1-p)^{m-T_m}$$

where $T_m = \sum_{i=1}^{m} X_i$ and is the total number of defectives among the m items drawn.

Now let

$$p_1(x) = p_1^x(1-p_1)^{1-x}$$

and

$$p_0(x) = p_0^x(1-p_0)^{1-x}$$

Referring to equations (9.12.1) and (9.12.4), we see that sampling continues as long as

$$\frac{\beta}{1-\alpha} < \frac{p_1(x_1) \times \cdots \times p_1(x_m)}{p_0(x_1) \times \cdots \times p_0(x_m)} < \frac{1-\beta}{\alpha}$$

that is, as long as

$$\frac{\beta}{1-\alpha} < \frac{p_1^{T_m}(1-p_1)^{m-T_m}}{p_0^{T_m}(1-p_0)^{m-T_m}} < \frac{1-\beta}{\alpha} \tag{9.12.8}$$

for $m = 1, 2, \ldots$.

Taking natural logarithms, we have that equation (9.12.8) is equivalent to:

$$\ln\frac{\beta}{1-\alpha} < T_m \ln\frac{p_1}{p_0} + (m - T_m)\ln\frac{1-p_1}{1-p_0} < \ln\frac{1-\beta}{\alpha}$$

If we plot the lines in the plane (m, T_m) given by

$$T_m \ln\frac{p_1}{p_0} + (m - T_m)\ln\frac{1-p_1}{1-p_0} = \ln\frac{\beta}{1-\alpha}$$

and

$$T_m \ln \frac{p_1}{p_0} + (m - T_m) \ln \frac{1-p_1}{1-p_0} < \ln \frac{1-\beta}{\alpha}$$

then as long as the sequence of sample points (m, T_m), $m = 1, 2, \ldots,$ stays inside the band formed by these two lines, we continue sampling. As soon as a point rises above the upper line, we reject H_0 in favor of H_1, or as soon as a point falls in the region below the lower line, we decide not to reject H_0.

9.12.2 A Two-Tail Sequential Testing Procedure

The one-sided sequential test can be easily adapted to the problem where we want to test the hypothesis H_0: $\mu = \mu_0$ against the two-sided alterative H_1: $\mu = \mu_0 + \delta$, where $\delta = \pm\delta_0$, say, with probabilities of type I and type II errors α and β, respectively. As in the fixed-sample-size test, it is customary to spread the α risk equally across both alternative hypotheses. The testing scheme then becomes two one-sided schemes with risks $\alpha/2$ and β, for the problems H_0: $\mu = \mu_0$ versus H_1: $\mu = \mu_0 + \delta_0$ and H_0: $\mu = \mu_0$ versus H_1: $\mu = \mu_0 - \delta_0$, which we would like to combine. Now, if we let $Y = X - \mu_0$, the two-sided sequential plot can be displayed in a single graph as illustrated in Figure 9.12.2. Employing equation

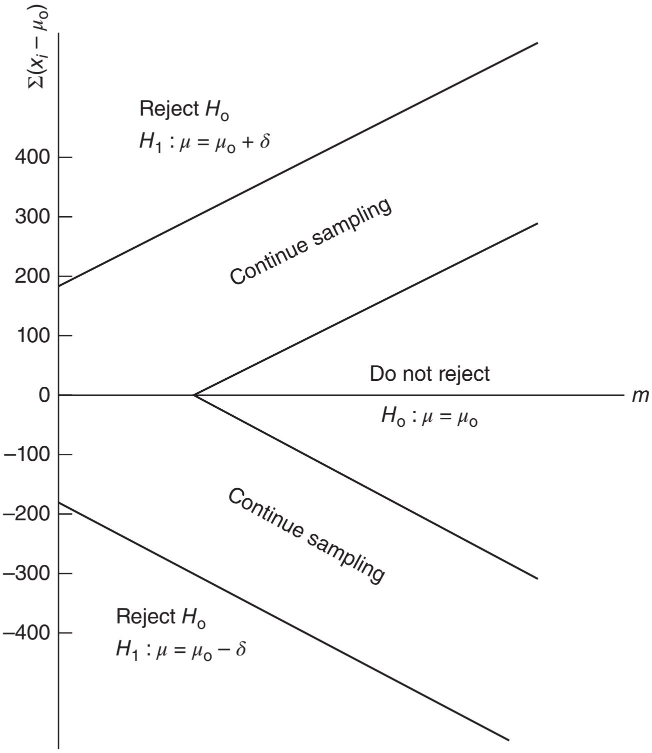


Figure 9.12.2 Graph of the two-sided sequential test of hypothesis for H_0: $\mu = \mu_0 = 500$; H_1: $\mu = \mu_0 \pm \delta$, $\delta = 50$, $\sigma = 50$, $\alpha = 0.05$, $\beta = 0.10$.

(9.12.5), we write the equations for the upper and lower pairs of boundary lines as

$$\left(\frac{\sigma^2}{\delta}\right)\ln\frac{\beta}{1-\alpha/2}+\frac{\delta}{2}m<\sum y_i<\left(\frac{\sigma^2}{\delta}\right)\ln\frac{1-\beta}{\alpha/2}+\frac{\delta}{2}m$$

where δ is the quantity specified in H_1 and $y_i = x_i - [(\mu_0 + \mu_1)/2]$.

Example 9.12.3 (A two-sided sequential testing procedure) *Suppose that* $\mu_0 = 500$, $\sigma = 50$, $\delta = \pm 50$, $\alpha = 0.05$, *and* $\beta = 0.10$. *Then, we obtain for the* upper pair *of boundary lines in Figure 9.12.2 (here* $T_m = \Sigma y_i$*)*

$$\begin{aligned} T_m &= 50(3.583) + 25m = 179.2 + 25m \\ T_m &= 50(-2.277) + 25m = -113.9 + 25m \end{aligned}$$

and for the lower pair *of boundary lines*

$$\begin{aligned} T_m &= -179.2 - 25m \\ T_m &= 113.9 - 25m \end{aligned}$$

As illustrated in Figure 9.12.2, so long as the cumulative sum $\Sigma(X_i - \mu_0)$ stays within the region interior to the two pairs of parallel control lines, another observation is taken. If the cumulative sum falls outside the boundaries of the outer lines, the hypothesis H_0: $\mu = \mu_0$ is rejected. The hypothesis H_0 is not rejected if the cumulative sum crosses into the region defined by the intersection of the boundaries that form the V-shaped region on the right in the figure.

PRACTICE PROBLEMS FOR SECTIONS 9.11 AND 9.12

1. A new and unusual plastic is to be produced by a pilot plant project using a standard extrusion method. Because of past experience with yield using a standard extrusion method, the yields, say X, are assumed to be normally distributed as $N(\mu, (20)^2)$, X measured in tons. The CEO hopes that $E(X) = \mu = 650$ tons. To test this, a sample of 50 days' production yields $\bar{X} = 630$ tons.
   (a) Find a 95% confidence interval for μ.
   (b) Use (a) to test the hypothesis H_0: $\mu = 650$ against H_1: $\mu \neq 650$ at the 5% level of significance.
2. Two plants are to produce certain fluorescent bulbs using new equipment. Because of the similarity to the other processes making fluorescent bulbs of different types, it is known that over wide ranges of μ_1 and μ_2, the distributions of the life of light bulbs from plants I and II are, respectively, $N(\mu_1, (200)^2)$ and $N(\mu_2, (200)^2)$. The quality control division at each of the two plants took data, and the results for the two random samples produced the following summary statistics:

$$\text{Plant I}: \ \bar{X}_1 = 1410 \text{ hours}, \ n_1 = 25$$

$$\text{Plant II}: \ \bar{X}_2 = 1260 \text{ hours}, \ n_2 = 20$$

   (a) Find a 99% confidence interval for $\delta = \mu_1 - \mu_2$.
   (b) Use the result in (a) to test the hypothesis H_0: $\delta = 0$ against H_1: $\delta \neq 0$ at the 1% level of significance.

3. A tire manufacturer has decided to bring out a new line of snow tires, equipped with studs that are inserted at the molding stage. The decision is made to test the amount of wear per 25,000 miles of use, and a standard test is planned. The weights before and after a tire is tested are recorded, and X, the loss in weight expressed as a percentage of initial weight for the tires, is then calculated. It is assumed that $X \sim N(\mu, \sigma^2)$. Four such tires, chosen at random and tested, yielded the following results:
$$n = 4, \quad \bar{X} = 19.4, \quad S^2 = 2.25$$
   (a) Find a 95% confidence interval for μ.
   (b) Use (a) to test the hypothesis H_0: $\mu = 18$ against H_1: $\mu \neq 18$ at the 5% level of significance.
4. Referring to Problem 3 previously mentioned, find a 99% confidence interval for $\sigma^2 = Var(X)$ and use it to test H_0: $\sigma^2 = 2$ versus H_1: $\sigma^2 \neq 2$.
5. Determinations of the percentage of chlorine in a batch of polymer are to be made by two analysts, 1 and 2, to see whether they do consistent work in this environment. Based on their experience working in this laboratory, it is assumed that determinations of these analysts are distributed as $N(\mu_i, \sigma^2)$, that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Ten determinations are made by each analyst, and the results are
$$\text{Analyst 1}: \quad \bar{X}_1 = 12.21, \ S_1^2 = 0.5419, \ n_1 = 10$$
$$\text{Analyst 2}: \quad \bar{X}_2 = 11.83, \ S_2^2 = 0.6065, \ n_2 = 10$$
   (a) Find a 95% confidence interval for $\mu_1 - \mu_2$.
   (b) Use the result in (a) to test the hypothesis H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 \neq 0$ at the 5% level of significance.
6. (a) In Problem 5 above find a 95% confidence interval for σ_1^2/σ_2^2 (i.e., do not assume that $\sigma_1^2 = \sigma_2^2$ in this problem).
   (b) Use part (a) to test the assumption in Problem 5 of equal variances at the 5% level of significance, that is test H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$.
7. A recently purchased rare coin is to be used to determine whether the new owner of the coin or his assistant will buy coffee at the morning office break. They agreed to keep records of the first 45 days' tosses to help decide whether the coin is fair. Indeed, the results are (X is the number of tosses that turn up heads):
$$n = 45, \quad X = 27$$
   (a) If p is the true probability that this new coin will show heads, establish a 95% confidence interval for p. (*Hint*: n is large.)
   (b) Use (a) to test the problem H_0: $p = 0.5$ versus H_1: $p \neq 0.5$
8. Generate sequentially random samples of size 5 from a binomial population with $n = 50$ and $p = 0.5$. Conduct a sequential test of the hypothesis H_0: $p = 0.5$ versus H_1: $p = p_1 > 0.50$ at significance level $\alpha = 0.05$ and power $1 - \beta = 0.90$ at $p = 0.52$; that is the desired value of the probability of type II error is to be $\beta = 0.10$ at $p = 0.52$. (*Hint*: A random sample can be generated using any one of the statistical packages discussed in this book. For example, MINITAB can generate a random sample from a given distribution as follows: select **Calc** > **Random data** > **Distribution,** and then select in the dialog box, the sample size and appropriate parameters of the distribution, and click **OK**.)

9. Generate sequentially random samples from a normal population with mean $\mu = 12$ and standard deviation $\sigma = 2$ to conduct a sequential test of the hypothesis $H_0\colon\ \mu = 20$ versus $H_1\colon\ \mu = \mu_1 > 20$ at significance level $\alpha = 0.05$ and power $1 - \beta = 0.95$ at $\mu = 21$; that is, the desired value of the probability of type II error is to be $\beta = 0.05$ at $\mu = 21$.
10. Generate sequentially random samples from an exponential distribution $f(x) = \lambda e^{-\lambda x}$ with $\lambda = 10$ to conduct a sequential test of the hypothesis $H_0\colon\ \lambda = 10$ versus $H_1\colon\ \lambda > 10$ at significance level $\alpha = 0.01$ and power $1 - \beta = 0.95$ at $\lambda = 12$; that is the desired value of the probability of type II error is to be $\beta = 0.05$ at $\lambda = 12$.

9.13 CASE STUDIES

Case Study 1 (Data source: a major integrated chip manufacturer) During the qualification of the product presented in the case study of Chapter 8, product **LMV9234**, the second lot was processed 30 days after the first lot. The polyresistor is a parameter for this product that is critical for the supply current and other parameters for this microchip. To produce a polyresistor, a polysilicon layer is provided with a boron implant. The first lot was processed on tool A and the second on tool B. Case Study 9.13.1 data on the book website provides polyresistor values. Determine 90%, 95%, and 99% confidence intervals for the difference of means of the polyresistor values for these lots. Further, the product engineer would like to compare (using 90%, 95%, and 99% confidence intervals) the means of the polyresistor values for these lots. Analyze the results of the case study. Prepare a short report summarizing your conclusions. The data for this case study are, as previously mentioned, available on the book website: www.wiley.com/college/gupta/statistics2e.

Case Study 2 (Data source: a major integrated chip manufacturer) As part of the final release for the LMV9234, ESD (electrostatic discharge), tests of the quality of this product were performed on 60 of the final packaged units. The ESD tests are stress tests that are performed for the human body model (HBM), machine model (MM), and charge device model (CDM). The LMV9234 is a 20-PIN microchip in an SOIC package (see Figure 9.13.1). Thirty units from one randomly selected wafer from lots 1 and 2 were built in the SOIC final package form and tested for HBM, CDM, and MM. The data for the units that passed for the HBM are listed on the book website, in Case Study 9.13.2 files. Passing units are designated by 1 and failing units by 0. Find 99% and 95% confidence

Figure 9.13.1 SOIC package.

intervals for the difference between the proportion of passing units between lots 1 and 2. Use these confidence intervals to test the hypothesis at the 1% and 5% level of significance that the proportions of passing units between lots 1 and 2 are the same. The data for this case study is available on the book website: www.wiley.com/college/gupta/statistics2e.

9.14 USING JMP

This section is not included in this book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Suppose that a certain type of 40-W bulbs is standardized so that its mean life is 1500 hours and the standard deviation is 200 hours. A random sample of 36 of these bulbs from lot L having mean μ was tested and found to have an average life of 1380 hours.
   (a) Test at the 1% significance level the hypothesis $H_0\colon\ \mu = 1500 \quad \text{versus} \quad H_1\colon\ \mu = \mu_1 < 1500$.
   (b) What is the power of the test at $\mu = 1400$?
   (c) Graph the power function.

2. Suppose that in Problem 2 of Section 9.3, we need to perform a test of $H_0\colon \mu = 0.25$ against $H_1\colon \mu = 0.2490$ with size $\alpha = 0.01$ and power at $\mu = 0.2490$ of 0.99, that is, $\beta = 0.01$. What sample size is necessary to achieve this?

3. Generalize the result in Problem 2. That is, suppose that the sampling is from $N(\mu, \sigma_0^2)$, where σ_0^2 is the known value of the variance of the population. Suppose that we need to test $H_0\colon \mu = \mu_0$ against $H_1\colon \mu = \mu_1$, where $\mu_1 < \mu_0$, so that the level of the test is α and $\gamma(\mu_1) = 1 - \beta$. Show that the sample size n used to achieve this is such that
$$\sqrt{n} = \frac{(z_\alpha + z_\beta)\sigma_0}{(\mu_0 - \mu_1)}.$$

4. Machines producing a particular brand of yarn are given periodic checks to help insure stable quality. A certain machine has been set in such a way that it is expected that strands of the yarn will have breaking strength of $\mu = 19.5$ oz, with a standard deviation of $\sigma = 1.80$ oz. (It has been found from experience that σ remains steady at 1.80 over a wide range of values of μ). A sample of 12 pieces of yarn selected randomly yields an average of breaking strengths of 18.46 oz. Assuming normality, test at the 5% level of significance the hypothesis $H_0\colon\ \mu = 19.50 \quad \text{versus} \quad H_1\colon\ \mu \neq 19.50$.

5. A liquor company is concerned about the filling of its bottles with “too little” or “too much” liquid. A filling machine set to fill “40-oz” bottles is checked by selecting at random 10 bottles and measuring their contents. Assuming that the machine fills the bottles with quantities distributed as $N(\mu,\ 1.44)$ for a wide range of values of μ, state the critical region for a 1% test of significance of $H_0\colon\ \mu = 40 \quad \text{versus} \quad H_1\colon\ \mu \neq 40$. Graph the power function.

6. A diamond-cutting machine has been set to turn out diamonds of “0.5-karat” weight. Assume from past experience that the machine produces diamonds of weight that

has the $N(\mu,\ 0.0036)$ distribution. It is important to a jewelry supply house that the weight not be too low (dissatisfied customers) or too high (economic considerations). Accordingly, the machine is checked every so often. A recent sample of six diamonds yielded weights of 0.48, 0.59, 0.54, 0.50, 0.55, and 0.56 karats. Test at the 5% level of significance the hypothesis $H_0\colon\ \mu = 0.5 \quad \text{versus} \quad H_1\colon\ \mu \neq 0.5$

7. A company is engaged in the stewed-fruit canning business. One of its brands is a medium-sized tin of cooked prunes that is advertised as containing 20 oz of prunes. The company must be sure that it packs prunes into the tins so that the mean weight is not "too much" under or over 20 oz (stiff fines would be imposed in the former case). A random sample of 14 cans yields an average of $\bar{X} = 20.82$ oz, with a standard deviation of 2.20 oz. Assuming that the weights of tins of prunes are distributed as $N(\mu, \sigma^2)$, test at the 5% level the hypothesis $H_0\colon\ \mu = 20 \quad \text{versus} \quad H_1\colon\ \mu \neq 20$

8. Five determinations of the percentage of nickel in a prepared batch of ore produced the following results:

   3.25, 3.27, 3.24, 3.26, 3.24

   If μ is the "true percentage of nickel in the batch", test the hypothesis $H_0\colon \mu = 3.25$ against the alternatives $H_1\colon \mu \neq 3.25$ at the 1% level of significance (assume normality).

9. Nine determinations were made by a technician of the melting point of manganese with the following results: 1268, 1271, 1259, 1266, 1257, 1263, 1272, 1260, 1256 in (degrees centigrade). Test at the 5% level of significance the hypothesis that the results are consistent with the published value of 1260°C for the true melting point of manganese (assume normality).

10. Suppose that a plot of land is surveyed by five student surveyors who find the following areas for the plot (in acres): 7.27, 7.24, 7.21, 7.28, 7.23. On the basis of this information, test the hypothesis that the true area of the plot is 7.23 acres or not at the 5% level of significance (assume normality).

11. A manufacturer claims that the diameters of rivets it produces have a standard deviation of $\sigma = 0.05$ inch. A sample of 16 rivets has a sample standard deviation of $S = 0.07$ in. Test at the 5% level of significance the hypothesis (assume normality): $H_0\colon\ \sigma = 0.05 \quad \text{versus} \quad H_1\colon\ \sigma > 0.05$

12. The standard deviation S of muzzle velocities of a random sample of 16 rounds of ammunition was found to be 85 ft/s. If the "standard" value of σ for the muzzle velocity of this type of ammunition is 78 ft/s, test at the 5% level of significance the hypothesis $H_0\colon \sigma = 78$ versus $H_1\colon \sigma > 78$ (assume normality).

13. Certain measurements were made on test pieces selected from two batches B_1 and B_2, with the following results shown:

Lot B_1	0.240	0.238	0.243	0.242	0.244	0.237
Lot B_2	0.185	0.190	0.192	0.186	0.188	0.190

   If μ_1 and μ_2 are the means and σ_1^2 and σ_2^2 are the variances of the measurements in batches B_1 and B_2, respectively, and assuming normality:

   (a) Test the hypothesis $H_0\colon \sigma_1^2/\sigma_2^2 = 1$ versus $H_1\colon \sigma_1^2/\sigma_2^2 \neq 1$ at the 5% level of significance.

(b) Using (a), test the hypothesis H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$ at the 5% level of significance.

14. Two methods were used in a study of the latent heat of fusing of ice. Both method A (an electrical method) and method B (a method of mixtures) were conducted with the specimens cooled to -0.72 °C. The data below represent the change in total heat from -0.72 °C, in calories per gram of mass:

Method A		Method B	
79.98	79.97	80.02	79.97
80.04	80.05	79.94	80.03
80.02	80.03	79.98	79.95
80.04	80.02	79.97	79.97
80.03	80.00		
80.04	80.02		
80.03			

Assuming normality of the population of determinations for each method as well as equal variances, test at the 5% level of significance the hypothesis that the change in total heat measured by method A is the same as that of method B.

15. Two methods for preparing fish, A and B, are compared according to a specific scoring scheme. The claim has been made that $\mu_A > \mu_B$. The composite scores of two samples are:

Method A		Method B	
4.05	4.18	3.31	2.35
5.04	4.35	3.39	2.59
3.45	3.88	2.24	4.48
3.57	3.02	3.93	3.93
4.23	4.56	3.37	3.43
4.23	4.37	3.21	3.13.

(a) Assuming normality of the scores generated by A and B, test at the 1% level of significance the hypothesis $\sigma_A^2 = \sigma_B^2$.
(b) Using the result in (a), test the hypothesis H_0: $\mu_A = \mu_B$ versus H_1: $\mu_A > \mu_B$.

16. Elongation measurements are made on 10 pieces of steel, five of which were treated with method A (aluminum plus calcium) and the remaining five with method B (aluminum only). It is conjectured that the addition of calcium will improve elongations by at least 1%. The results of the measurements are (in percent):

Method A	Method B
34	28
27	29
30	25
26	23
33	30.

Assuming normality:

(a) Test the hypothesis $\sigma_A^2 = \sigma_B^2$ at the 5% level of significance.

(b) Using the result in (a), test the hypothesis H_0: $\mu_A - \mu_B = 1\%$ versus H_1: $\mu_A - \mu_B > 1\%$

17. A comparison of yields of marigolds from control plots and treated plots is carried out. Samples from eight control plots and eight treated plots yield the following data:

$$\text{Treated } (A): \quad n_A = 8, \quad \bar{x}_A = 128.4, \quad s_A^2 = 117.1$$

$$\text{Nottreated } (B): \quad n_B = 8, \quad \bar{x}_B = 96.5, \quad s_B^2 = 227.7$$

Assuming normality:

(a) Test at the 1% level of significance $\sigma_A^2 = \sigma_B^2$.
(b) Using the result in (a), test the hypothesis H_0: $\mu_A - \mu_B = 0$ versus H_1: $\mu_A - \mu_B > 0$.

18. Viewing times of members of households in two different types of communities are sampled, with the following results:

$$\text{Community 1}: \; n_1 = 40, \; \bar{X}_1 = 19.2 \text{ hours/week}, \; S_1^2 = 6.4$$

$$\text{Community 2}: \; n_2 = 50, \; \bar{X}_2 = 15.9 \text{ hours/week}, \; S_2^2 = 3.2$$

(a) Assuming normality, test the hypothesis $\sigma_1^2 = \sigma_2^2$ at the 1% level of significance.
(b) Using the result in (a), test the hypothesis H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 > \mu_2$.

19. It has been suspected for some time that the morning shift is more efficient than the afternoon shift. Random observations yield the following data:

$$\text{Morning shift}: \; n_1 = 5, \; \bar{X}_1 = 22.9, \; S_1^2 = 6.75$$
$$\text{Afternoon shift}: \; n_2 = 7, \; \bar{X}_2 = 21.5, \; S_2^2 = 7.25$$

Assuming normality:

(a) Test the hypothesis $\sigma_1^2 = \sigma_2^2$ at the 1% level of significance.
(b) Using (a), test the hypothesis H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 > \mu_2$.

20. A sample of 220 items turned out (during a given week by a certain process) to have average weight of 2.46 lb and standard deviation of 0.57 lb. During the next week, a different lot of raw material was used, and the average weight of a sample of 205 items turned out that week to be 2.55 lb and the standard deviation was 0.48 lb. Assuming normality and equality of variances, would you conclude from these results that the mean weight of the product had increased significantly at the 5% level of significance during the second week?

21. Orange juice cans are filled using two methods. Two random samples one from each method produced the following results:

$$\text{Method 1}: \; n_1 = 40, \; \bar{X}_1 = 21.78, \; S_1^2 = 3.11$$

$$\text{Method 2}: \; n_2 = 40, \; \bar{X}_2 = 20.71, \; S_2^2 = 2.40$$

(a) Assuming normality, test at the 5% level of significance H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$.
(b) Test, using the result of (a), the hypothesis H_0: $\mu_1 = \mu_2$, versus H_1: $\mu_1 > \mu_2$.

22. The systolic blood pressure of a group of 70 patients yielded $\bar{X}_1 = 145$ and $S_1 = 14$. A second group of 70 patients, after being given a certain drug, yielded $\bar{X}_2 = 140$ and $S_2 = 9$.

(a) Assuming normality, test at the 5% level of significance H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$.
(b) Using the result of (a), test at the 5% level of significance the hypothesis H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 > \mu_2$.

23. Two randomly selected groups of 70 trainees each are taught a new assembly line operation by two different methods, with the following results when the groups are tested:

$$\text{Group 1}: \; n_1 = 70, \; \bar{X}_1 = 268.8, \; S_1 = 20.2$$

$$\text{Group 2}: \; n_2 = 70, \; \bar{X}_2 = 255.4, \; S_2 = 26.8$$

(a) Assuming normality, test the hypothesis $\sigma_1 = \sigma_2$ versus $\sigma_2 > \sigma_1$. Use $\alpha = 0.01$.
(b) Using the result of (a), test at the 5% level of significance the hypothesis H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 > \mu_2$.

24. The following data give the results for iron content of ore using two methods A (dichromate) or B (thioglycolate) on 10 different samples of ore.

Sample number	Percent iron by method A	Percent iron by method B
1	28.22	28.27
2	33.95	33.99
3	38.25	38.20
4	42.52	42.42
5	37.62	37.64
6	37.84	37.85
7	36.12	36.21
8	35.11	35.20
9	34.45	34.40
10	52.83	52.86

Assuming normality, use the method of paired comparisons to test whether these two methods yield significantly different percentages of iron at the 5% level of significance.

25. Analysts I and II each make a determination of the melting point in degrees centigrade of hydroquinine on each of eight specimens of hydroquinine with the results shown below:

Specimen number	Analyst I (°C)	Analyst II (°C)
1	174.0	173.0
2	173.5	173.0
3	173.0	172.0
4	173.5	173.0
5	171.5	171.0
6	172.5	172.0
7	173.5	171.0
8	173.5	172.0

Using the method of paired comparison, do the methods of determinations differ significantly at the 5% level of significance? (Assume normality.)

26. Over a long period of time, 10 patients selected at random are given two treatments for a specific form of arthritis. The results (in coded units) are given below:

Patients	Treatment 1	Treatment 2
1	47	52
2	38	35
3	50	52
4	33	35
5	47	46
6	23	27
7	40	45
8	42	41
9	15	17
10	36	41

Is there a difference in efficacy of the two treatments? Use $\alpha = 0.05$ and assume normality.

27. Two different methods of storing chicken are contrasted by applying technique 1 (a freezing technique) to one-half of a chicken and technique 2 (a wrapping technique) to the other half of the same chicken. Both halves are stored for three weeks, and a certain "tenderness of the meat" test is then applied. This is done for 200 chickens, and using the notation of paired comparison, it was found that $\bar{X}_d = -2.430$ with $S_d = 0.925$. Test the hypothesis, at the 5% level of significance, that $\mu_d = -1$ versus $\mu_d \neq -1$, where $\mu_d = E(Y_2) - E(Y_1)$.

28. The dependability of analysts is occasionally measured by the variability of their work. Two analysts A and B each make 10 determinations of percent of iron content in a batch of prepared ore from a certain deposit. The sample variances obtained are $S_A^2 = 0.4322$ and $S_B^2 = 0.5006$. Are the analysts equally dependable? Test at the 5% level of significance and assume normality.

29. In Problem 2 of Section 9.7, it was assumed that $\sigma_A^2 = \sigma_B^2$. Test this assumption at the 1% level of significance.

30. In Problem 14, is the assumption that $\sigma_A^2 = \sigma_B^2$ warranted on the basis of the data? Use the significance level of 0.05.

31. In Problem 20, is the assumption of equality of variances valid? Use the significance level 0.01.

32. Using the data of Problem 24, test at the 1% level of significance the hypothesis $\sigma_A^2 = \sigma_B^2$.

33. Using the data of Problem 25, test at the 1% level of significance the hypothesis $\sigma_I^2 = \sigma_{II}^2$.

34. If X is a random variable with probability function

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

describe with $\alpha = 0.05$, $\beta = 0.01$, how to test sequentially the hypothesis H_0: $\lambda = \lambda_0 = 1.5$ versus H_1: $\lambda = \lambda_1 = 2.0$.

35. If X is a random variable with probability function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

describe, with $\alpha = 0.10$, $\beta = 0.05$, how to test sequentially the hypothesis H_0: $p = 0.10$ versus H_1: $p = 0.20$.

36. If a random variable X has probability density function $f(x)$ given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2\sigma^2)(x-\mu_0)^2}$$

where μ_0 is the known value of the population mean, describe with $\alpha = 0.10$, $\beta = 0.05$, how sequentially to test the hypothesis H_0: $\sigma^2 = \sigma_0^2$ versus H_1: $\sigma^2 = \sigma_1^2 > \sigma_0^2$.

37. An advising committee at a university is interested in finding the mean time spent by all the university students in watching sports on television. The time spent in watching sports on television by a random sample of 49 students produced a sample mean $\bar{X}$ of 7.5 hours and a standard deviation S of 1.4 hours. Assuming normality test, the hypothesis H_0: $\mu = 7.0$ against H_1: $\mu > 7.0$ at the 1% level of significance.

38. The mean and the standard deviation of daily intake of vitamin D for a random sample of 36 girls between the ages 16 and 20 years is 450 and 50 mg, respectively. Assuming normality, test the following at the 5% level of significance:
    (a) H_0: $\mu = 500$ against H_1: $\mu < 500$
    (b) H_0: $\mu = 500$ against H_1: $\mu \neq 500$

39. In Problem 38, find the p-value for the hypotheses in (a) and (b).

40. The lifetime (in hours) of AAA batteries used in TI-83 series calculators is assumed to be normally distributed. A random sample of 100 such batteries produced a sample mean of 58.7 hours and a sample standard deviation of 2.5 hours. Test at the 1% level of significance the hypotheses:
    (a) H_0: $\mu = 60$ versus H_1: $\mu < 60$
    (b) H_0: $\mu = 60$ versus H_1: $\mu \neq 60$
    Find the p-values in (a) and (b).

41. A rod used in auto engines is required to have a diameter of 18 millimeters (mm). A random sample of 64 rods produced a sample mean of 18.2 mm and a standard deviation of 1.2 mm. Assuming normality, test the hypothesis H_0: $\mu = 18$ versus H_1: $\mu \neq 18$ at the 5% level of significance. Find the p-value. Find the power of the test if the true mean of the manufactured rods is 18.5 mm and $\sigma = 1.2$ mm.

42. The piston rings of certain diameter in mm for automobile engines are being manufactured at two plants. Two random samples of piston rings, one sample from each

plant, are taken, and the diameters of the rings are measured. These data produce the following sample statistics:

$$n_1 = 50, \quad \bar{X}_1 = 73.54, \quad S_1 = 0.2; \quad n_2 = 50, \quad \bar{X}_2 = 74.29, \quad S_2 = 0.15$$

Assuming normality:

(a) Test at 5% level of significance test the hypothesis $H_o : \mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$.
(b) Find the observed level of significance (p-value).
(c) Using the p-value found in (b), decide whether or not the null hypothesis should be rejected.

43. A chemical manufacturing company is interested in increasing the yield of a certain chemical. To achieve this goal, a chemical engineer decides to use the catalyst at two different temperatures, 300 and 350 °C. Two samples, each of size $n = 49$, are produced at each of these temperatures, and the output of the chemical is measured in grams. These data produce the following statistics:

$$\bar{X}_1 = 68.8, \quad S_1 = 5.1; \quad \bar{X}_2 = 81.5, \quad S_2 = 7.4$$

Assuming normality:

(a) Test at the 5% level of significance hypothesis H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 < 0$.
(b) Find β the probability of the type II error, and the power of the test if the true difference is 5 grams and $\sigma_1 = 5.1, \sigma_2 = 7.4$.

44. A manager of a large bank wants to compare the loan amounts of two of her loan officers. The loans issued by the two sales managers during the past three months furnished the following summary statistics:

$$n_1 = 55, \quad \bar{X}_1 = \$68{,}750, \quad S_1 = \$4{,}930; \quad n_2 = 60, \quad \bar{X}_2 = \$74{,}350, \quad S_2 = \$5{,}400$$

Assuming normality:

(a) Do these data provide sufficient evidence to indicate that the two loan officers issue loans of equal value? Use $\alpha = 0.05$.
(b) Do these data provide sufficient evidence to indicate that the loans by officer one are less than those of the other officer? Use $\alpha = 0.01$.

45. A random sample of 64 cigarettes of a particular brand yielded mean tar content per cigarette of 15.5 milligrams (mg) and a standard deviation of 1.4 mg.
Assuming normality:

(a) Test a hypothesis H_0: $\mu = 15$ versus H_1: $\mu > 15$ at the 0.01 level of significance.
(b) Find β the probability of the type II error, and the power of the test if the true value of the mean tar content per cigarette is 16 mg.

46. Observations of a random sample from a normal population with unknown mean μ and unknown standard deviation σ are

25	20	23	28	26	21	30	29	23	29

(a) Test at the 1% level of significance H_0: $\mu = 25$ versus H_1: $\mu \neq 25$.
(b) Find the p-value for the test in (a).

47. A health insurance company wants to find the average amounts of benefits it pays for a typically insured family of four (a couple with two children). The company selected a random sample of 16 such families and found that it paid on average $4858 with a standard deviation of $575. Assuming normality, test at the 5% level of significance the hypothesis H_0: $\mu = 5000$ versus H_1: $\mu < 5000$.Find the p-value for this test.

48. In a study of pregnancy-induced hypertension, two randomly selected groups of women with this diagnosis were selected. One group was treated for a certain period with an aspirin-based medicine and the other group was given a placebo. After the period of treatment, their arterial blood pressures in millimeters of mercury (mmHg) were checked. The study produced the following test statistics:

$$n_1 = 12, \;\; \bar{X}_1 = 110\,\text{mmHg}, \;\; S_1 = 9\,\text{mmHg};$$

$$n_2 = 14, \;\; \bar{X}_2 = 115\,\text{mmHg}, \;\; S_2 = 10\,\text{mmHg}$$

Assuming that the two populations are normally distributed with equal variances, test at the 1% level of significance the hypothesis $H_o : \mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$. Find the p-value for the test.

49. Repeat Problem 48, assuming that the population variances are not equal.

50. Two random samples from two normal populations with standard deviations $\sigma_1 = 4.5$ and $\sigma_2 = 6.2$, respectively, produced the following data:

Sample from population I	20	30	31	28	34	35	32	26	24	38	25	40
Sample from population II	34	36	49	52	41	44	30	33	47	49	39	

(a) Test at the 2% level of significance the hypothesis $H_o : \mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$.
(b) Find the p-value for the test and make your conclusions using the p-value. Do you arrive at the same conclusion as in (a)?

51. A medical school professor claims that medical students study (including the time in class) on average at least 16 hours a day. To verify the professor's claim, a random sample of 19 students was taken, and each student was asked about the number of hours he or she spends studying each day. This inquiry resulted in the following data:

15	23	18	16	17	15	16	14	17	18	17	15	14	16	15	17	13	15	14

(a) Formulate an appropriate hypothesis to test the professor's claim.
(b) Test the hypothesis you formulated in (a), using $\alpha = 0.05$. Assume normality.
(c) Find the p-value for the test.

52. The following are the numbers of defective parts produced in a shift by 10 workers before and after going through a very rigorous training program (assuming normality):

Worker	1	2	3	4	5	6	7	8	9	10
Before	15	12	16	14	18	12	13	17	10	16
After	8	5	10	5	14	4	6	6	3	12

(a) Do these data provide sufficient evidence to indicate that the training program was effective? Use $\alpha = 0.05$.

(b) Find the p-value for the test.

53. The following data shows the weight gain (in pounds) in one week for a sample of 10 pigs before and after they were given a type of hormone:

Pig number	1	2	3	4	5	6	7	8	9	10
Before	10	8	9	11	7	8	6	12	8	9
After	15	13	12	10	11	9	11	15	12	16

Assuming normality:

(a) Formulate a hypothesis to test the effectiveness of hormones in inducing weight gain.

(b) Test the hypothesis in (a) at the 1% level of significance.

54. An owner of two stores wants to evaluate the customer service at his/her stores. In order to do this, the owner took a random sample of 400 customers from store I and of 500 customers from store II. He asked the customers to rate the service at the store as excellent or not excellent and found that 250 of the 400 and 380 of the 500 customers rated the service as excellent.

(a) Test at the 5% level of significance that the proportions of customers at the two stores who think that the service is excellent is the same in both stores against the proportion of customers who think that service is excellent in store I, but lower than that in store II.

(b) Test at the 2% level of significance that the proportions of customers who thinks the service is excellent at the two stores are the same.

55. A patron of a casino doubts that the dice used in the casino are balanced. During a visit, she rolled a die 100 times and got an even number only 30 times.

(a) Formulate the hypothesis you would use to test whether the die is fair.

(b) At what observed level of significance would this null hypothesis be rejected?

56. A manufacturer of brass bolts has two plants. A random sample of 300 bolts from plant I showed that 21 of them were defective. Another random sample of 425 bolts from plant II showed that 24 of them were defective. Testing at the 5% level of significance, can you conclude that the proportions of defective bolts at the two plants are the same? Find the p-value for the test.

57. A six-sigma black belt quality control engineer found that in a random sample of 140 printed circuit boards, 18 are defective due to the result of certain nonconformity tests. At the 5% level of significance, test that the percentage of defective printed circuit boards is 10% against the alternative that it is greater than 10%. Find the p-value for the test.

58. A random sample of 18 observations from a normal population produced a sample mean of 37.4 and a sample variance of 15.6. Do the data provide sufficient evidence to indicate that $\sigma^2 < 20$? Use the 5% level of significance.

59. A machine is calibrated to fill bottles with 16 oz of orange juice. A random sample of 12 bottles was selected, and the actual amount of orange juice in each bottle was measured. The data are as follows:

15.0	15.9	15.4	16.1	15.2	15.8	16.4	15.7	15.8	16.3	16.5	16.2

Assuming normality, test at the 1% level of significance the hypothesis $H_0 : \sigma^2 = 0.2$ versus $H_1 : \sigma^2 \neq 0.2$.

60. An endocrinologist measured the serum levels of lipid peroxides (LP) among subjects with type I diabetes and also among normal subjects. These data produced the following summary statistics.

$$\text{Diabetic subjects}: \; n_1 = 25, \; \bar{X}_1 = 2.55, \; S_1^2 = 1.475$$

$$\text{Normal subjects}: \; n_2 = 36, \; \bar{X}_2 = 2.25, \; S_2^2 = 0.878$$

Test at the 5% level of significance the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ versus H_0: $\sigma_1^2 \neq \sigma_2^2$. Assume that the LP levels in two groups are normally distributed with variances σ_1^2 and σ_2^2, respectively.

61. In Problem 44, test at the 5% level of significance the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 < \sigma_2^2$

62. Pull-strength tests were applied on soldered lead in an electronic apparatus for each of two independent random samples of 12. The lead soldering in the two samples were done using two different techniques. The test results indicate the force required in pounds to break the bond. The data obtained from these two experiments are as follows:

Apparatus	1	2	3	4	5	6	7	8	9	10	11	12
Sample I	21.5	20.6	18.7	22.3	24.1	20.6	19.8	18.7	24.2	22.3	19.5	20.6
Sample II	24.6	23.5	22.5	23.5	22.7	21.5	20.5	23.6	22.5	23.5	21.7	19.9

Assuming that the pull-strengths are normally distributed with variances σ_1^2 and σ_2^2, respectively, test at the 5% level of significance the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$.

63. The monthly returns in percentage of dollars of two investment portfolios were recorded for one year. The results obtained are

Portfolio I	2.1	1.2	−1.5	1.9	0.7	2.5	3.0	−2.2	1.8	0.5	2.0	1.5
Portfolio II	2.9	3.5	−2.8	1.0	−3.0	2.6	−3.5	4.5	1.5	2.3	−1.0	0.8

Assume that the monthly returns of two portfolios are normally distributed with variances σ_1^2 and σ_2^2, respectively, and test at the 5% level of significance the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ versus H_0: $\sigma_1^2 \neq \sigma_2^2$.

64. Referring to Problem 1 of Section 8.7, test at the 5% level of significance the hypothesis H_0: $p = 0.45$ versus H_1: $p < 0.45$.

65. Use the confidence interval you obtained in Problem 1 of Section 8.7 to test, at the 5% level of significance, the hypothesis H_0: $p = 0.45$ versus H_1: $p \neq 0.45$.

66. Referring to Problem 2 of Section 8.7, test at the 5% level of significance the hypothesis that the coin is unbiased, that is, H_0: $p = 0.5$ versus H_1: $p \neq 0.5$.

67. Referring to Problem 3 of Section 8.7, test at the 5% level of significance the hypothesis that the proportions of the population favoring brand X before and after the advertising campaign are the same, that is, H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 \neq 0$.

68. Use the 99% confidence interval you obtained in Problem 47 of the Review Practice Problems in Chapter 8 to test at the 1% level of significance the hypothesis H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 \neq 0$.

69. Referring to Problem 5 of Section 8.7, test at the 5% level of significance the hypothesis that the percentages of persons who favor a nuclear plant in the two states are the same, that is, H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 \neq 0$.

70. Referring to Problem 5 of Section 8.7, determine a 98% confidence interval for the difference between the percentages of persons who favor a nuclear plant in their state and then use it to test at the 2% level of significance the hypothesis H_0: $p_1 - p_2 = 0$ versus H_1: $p_1 - p_2 \neq 0$.

Part II

Statistics in Actions

Chapter 10

ELEMENTS OF RELIABILITY THEORY

The focus of this chapter is a discussion of the basic concepts of reliability theory.

Topics Covered

- Reliability and hazard rate functions
- Derivation of an appropriate form for the life distribution using the hazard rate function
- Estimation and testing of hypotheses for parameters of different life distributions
- Probability plots and plots of the life distribution and hazard rate functions using MINITAB, R, and JMP for censored and uncensored data

Learning Outcomes

After studying this chapter, the reader will be able to

- Determine the reliability and hazard rate functions for different life probability models.
- Fit different models such as exponential, normal, lognormal, and Weibull to a censored or uncensored data set.
- Determine estimates of parameters of different life probability models.
- Perform testing of hypotheses concerning various parameters of life probability models.
- Use various statistical packages to construct survival plots using censored and uncensored data.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

10.1 THE RELIABILITY FUNCTION

Let T represent the random variable which is the time to failure of an item and let $f(t)$ be its probability density function (p.d.f.), where $T \geq 0$. Then, we define the *reliability function* $R(t)$ at time t as

$$R(t) = P(T \geq t) = \int_t^\infty f(x)dx = 1 - F(t) \tag{10.1.1}$$

It follows from equation (10.1.1) that $R(0) = 1$ and $R(\infty) = 0$. Note that the reliability function $R(t)$ is a nonincreasing function; that is the chances of survival diminish as the component becomes older. Here, $F(t)$ is the cumulative distribution function (c.d.f.) of the random variable T, the time to failure. For example, suppose the time to failure T is characterized by the exponential distribution; that is the process under investigation follows a Poisson process. Then, the reliability $R(t)$, for $t > 0$ is given by

$$R(t) = \int_t^\infty \lambda e^{-\lambda x}dx = e^{-\lambda t} \tag{10.1.1a}$$

where $1/\lambda$ is the mean of the exponential distribution. Figure 10.1.1 shows graphs of a typical density function, distribution function, and reliability function.

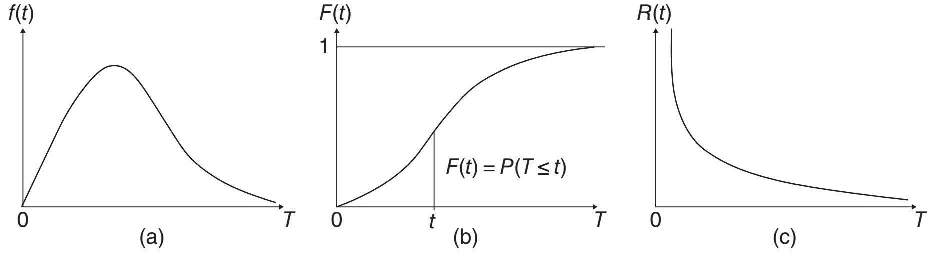


Figure 10.1.1 Plots of a typical (a) density function, (b) distribution function, and (c) reliability function.

10.1.1 The Hazard Rate Function

To assist in discriminating among distributions for time to failure, the concept of the *hazard rate function* or *failure rate function*, $h(t)$, plays an important role. We now define the hazard or failure rate function, denoted by $h(t)$, as the conditional probability that a system will fail almost instantaneously after time t, given that it has survived up to time t, that is

$$h(t) = \lim \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t} \text{ as } \delta t \to 0 \tag{10.1.2}$$

Now

$$P(t \leq T \leq t + \delta t | T \geq t) = \frac{P(t \leq T \leq t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{1 - F(t)}$$

and it follows that the hazard function may also be defined in terms of the density function and the reliability function as

$$\boxed{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)} \qquad (10.1.3)}$$

Note that the hazard rate may be increasing if aging has a deleterious effect, decreasing if aging is beneficial, constant, or even nonmonotone. For example, the Weibull distribution (defined in equation (10.1.9)) has an increasing, decreasing, or constant hazard rate depending on whether the shape parameter is greater than one, less than one, or equal to one (i.e., aging has no effect on survival). The normal distribution has an increasing hazard rate while, for example, the Pareto distribution has a decreasing hazard rate. These examples show that the hazard rate characterizes the aging process of the component.

The hazard rate is often called the *instantaneous failure rate*, and the *force of mortality* by actuaries. To explain, consider a young person and an old person, each with identical value for $f(t)$ at this very moment; that is both individuals have the same probability of dying in the next instant. The force of mortality $h(t)$ is greater for the older person since his or her $R(t) = 1 - F(t)$ is much less than that of the younger person.

Example 10.1.1 (Determining hazard function) *Suppose that the time to failure of light bulbs is distributed by the $N(\mu, \sigma^2)$ distribution with $\mu = 1600$ hours and $\sigma = 400$ hours. Thus, the reliability and the hazard rate at $t = 1700$ hours are*

$$R(t = 1700) = 1 - \Phi\left(\frac{1700 - 1600}{400}\right) = 1 - 0.5987 = 0.4013$$

Here, from equation (10.1.3), the hazard at $t = 1700$ is

$$h(t = 1700) = \frac{f(1700)}{R(1700)}$$

where $f(t)$ is the ordinate of the p.d.f. of the $N(1600, 400^2)$ distribution, that is

$$f(1700) = (2\pi \times 400^2)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{1700 - 1600}{400}\right)^2\right] = 9.6648 \times 10^{-4}$$

The instantaneous failure rate at $t = 1700$ is

$$h(t) = \frac{9.6648 \times 10^{-4}}{0.4013} = 24.08 \times 10^{-4}$$

Note that as stated earlier, the hazard rate function can be increasing, decreasing, or constant. When the hazard rate is constant, the distribution describing time to failure is the *exponential distribution*. When the hazard rate is increasing or decreasing, this behavior implies that the aging/antiaging factor is being taken into account, and the distribution describing time to failure is the Weibull distribution, gamma, or some other distribution. For example, a hazard rate function for an aging system that is not well maintained may increase or decrease after it is reconditioned and remain constant if it is updated or maintained frequently. In human populations, the hazard rate function is defined as the proportion of individuals alive at age t years who will die in the age interval $(t, t + \delta t)$. The failure rate curve for humans is almost bathtub-shaped, initially decreasing, then almost flat, and finally increasing. Figure 10.1.2 shows four kinds of hazard rate functions, $h_1(t), h_2(t), h_3(t)$, and $h_4(t)$, that are respectively increasing, decreasing, constant, and bathtub-shaped.

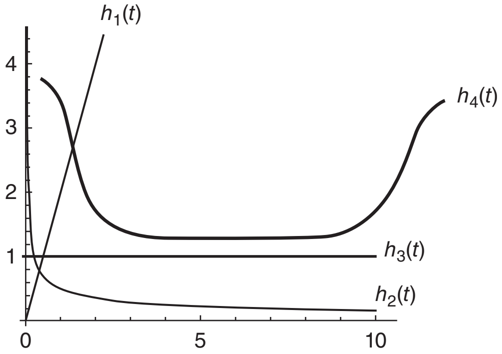


Figure 10.1.2 Graphs of four different hazard rate functions.

Example 10.1.2 (Microprocessor failure rate) *The time T to failure of a microprocessor is exponentially distributed with mean time to failure $\mu = 1/\lambda = 2000$ hours. It is important that the microprocessors work successfully over the first 200 hours of use. Find the reliability at $T = t$, where $t = 200$.*

Solution: Using equation (10.1.1a), we find that the reliability at $t = 200$ as

$$R(t = 200) = e^{-200(1/2000)} = 0.9048$$

Thus, 90.48% of the microprocessors will operate over the intended 200-hour period.

(In biometry, the reliability function $R(t)$ is often termed the survival function at time t.) Further, in this example, using equation (10.1.3), the instantaneous failure rate at $t = 200$ is

$$h(200) = \frac{f(200)}{R(200)} = \frac{(1/2000)e^{-200/2000}}{e^{-200/2000}}$$

that is

$$h(200) = \lambda = 1/2000$$

There are many devices for which a constant hazard rate is appropriate. Consider a very large sample of items all placed in operation at time $t = 0$. Substandard items will quickly fail, and as other weaker items are eliminated, it is common to see the hazard rate fall rapidly until it reaches a near-constant value.

If the random variable T, the time to failure of a system, is distributed as *gamma*, then from equation (5.9.10), we have

$$f(t|\gamma,\lambda)=\frac{\lambda^{\gamma}}{\Gamma(\gamma)}t^{\gamma-1}e^{-\lambda t},\quad t\geq 0;\gamma>0,\lambda>0 \tag{10.1.4}$$

Thus, the c.d.f. $F(t)$ is given by

$$F(t)=\frac{\lambda^{\gamma}}{\Gamma(\gamma)}\int_0^t x^{\gamma-1}e^{-\lambda x}dx \tag{10.1.5}$$

Let $y=\lambda x$; we have

$$F(t)=\frac{1}{\Gamma(\gamma)}\int_0^{\lambda t} y^{\gamma-1}e^{-y}dy \tag{10.1.6}$$

which is an incomplete gamma function, denoted by $I(\gamma,\lambda t)$. Thus, the reliability function $R(t)$ is given by

$$R(t)=1-F(t)=1-I(\gamma,\lambda t) \tag{10.1.7}$$

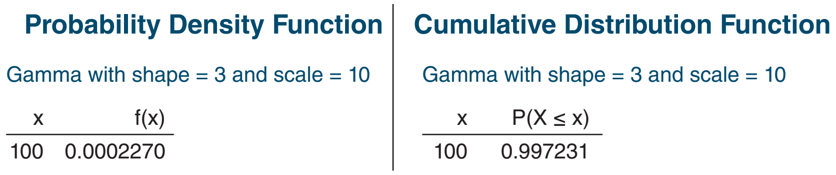
The values of the reliability function $R(t)$ can easily be determined by one of the statistical software packages used in this book. Using equations (10.1.2) and (10.1.7), the hazard rate function is given by

$$h(t)=\frac{f(t)}{1-I(\gamma,\lambda t)} \tag{10.1.8}$$

Example 10.1.3 (Microprocessor failure rate) *The time T (in hours) to failure of a microprocessor is modeled by a gamma distribution with parameters $\gamma=3$ and $\lambda=0.1$, so that the shape parameter is 3 and the scale parameter is 10. Using MINITAB and R, find the reliability and hazard rate function at $T=t$, where $t=100$ (see Figure 10.1.3 for the shape of $h(t)$ when $\gamma=3$, and $\lambda=1.0$ and 0.5 and for $\gamma=1$ and $\lambda=1.0$ and 0.5).*

MINITAB

Using Minitab (see Chapter 5), we have



```
Probability Density Function          | Cumulative Distribution Function
                                      |
Gamma with shape = 3 and scale = 10   | Gamma with shape = 3 and scale = 10
                                      |
  x        f(x)                       |   x      P(X ≤ x)
100   0.0002270                       | 100      0.997231
```

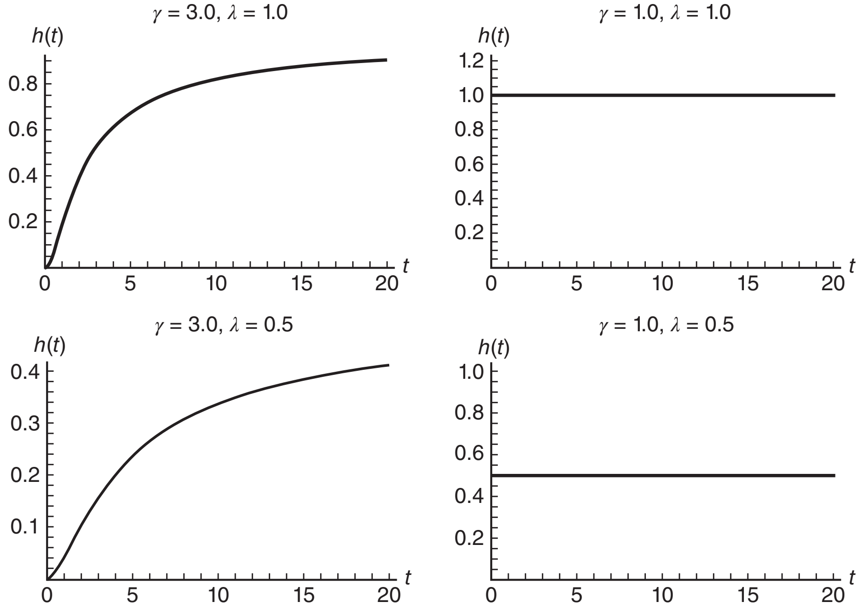


Figure 10.1.3 Graphs of the hazard function for the gamma distribution at various values of γ and λ.

Thus, the reliability at 100 is given by

$$R(100) = 1 - F(100) = 0.002769$$

The hazard rate function at $t = 100$ is given by $h(100) = f(100)/R(100) = 0.000227/0.002769 = 0.0819$.

USING R

R has a built-in gamma density function ‘dgamma(x, shape, rate, scale = 1/rate)’ and a gamma c.d.f. ‘pgamma(q, shape, rate, scale = 1/rate)’, where both ‘x’ and ‘q’ represent the quantile, and ‘shape’ and ‘scale’ are the shape and scale parameters of the gamma distribution, respectively. Alternatively, one can specify the rate parameter, which is equal to 1/scale. Referring to Example 10.1.3, first we should find density and distribution function values at $t = 100$ and use those results to evaluate the reliability and hazard rate functions.

```
f.100 = dgamma(x = 100, shape = 3, scale = 10)
F.100 = pgamma(q = 100, shape = 3, scale = 10)


f.100
[1] 0.0002269996
F.100
[1] 0.9972306
```

```
#The reliability at t = 100
R.100 = 1-F.100
R.100 #R output
[1] 0.002769396


#The hazard rate function at t = 100
h.100 = f.100/R.100
h.100 #R output
[1] 0.08196721
```

It is important to note here that the hazard rate function of the gamma distribution is increasing if $\gamma > 1$, decreasing if $\gamma < 1$, and constant if $\gamma = 1$ (see Figure 10.1.3).

Consider now the case where the random variable T is distributed as *Weibull* so that from equations (5.9.18) and (5.9.19), we have

$$f(t|\alpha, \beta, \tau) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{t-\tau}{\alpha}\right)^{\beta-1} e^{-[(t-\tau)/\alpha]^{\beta}}, & \text{for } t > \tau \\ 0, & \text{otherwise} \end{cases} \tag{10.1.9}$$

where $\alpha > 0, \beta > 0, \tau \geq 0$, are the parameters of the distribution. Here, α is called as the *scale parameter*, β the *shape parameter*, and τ the *location* or *threshold parameter*.

The c.d.f. of the Weibull distribution is given by

$$F(t) = \begin{cases} P(T \leq t) = 1 - e^{-[(t-\tau)/\alpha]^{\beta}}, & \text{for } t > \tau \\ 0, & \text{for } t < \tau \end{cases} \tag{10.1.10}$$

We use equation (10.1.10) to obtain the reliability and hazard rate functions:

$$R(t) = 1 - P(T \leq t) = 1 - (1 - e^{-[(t-\tau)/\alpha]^{\beta}}) = e^{-[(t-\tau)/\alpha]^{\beta}} \tag{10.1.11}$$

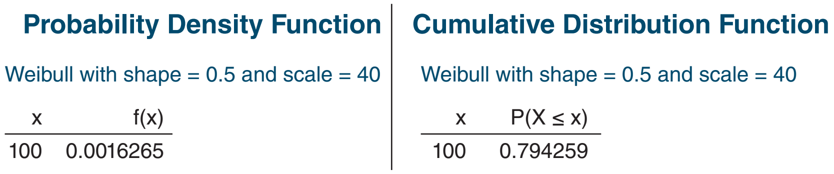
$$h(t) = \frac{\beta}{\alpha}\left(\frac{t-\tau}{\alpha}\right)^{\beta-1} \tag{10.1.12}$$

respectively. From equation (10.1.12), it can be seen that the hazard rate function of the Weibull distribution is increasing if $\beta > 1$, decreasing if $\beta < 1$, and constant if $\beta = 1$. Thus, the shape parameter determines the monotonicity of the hazard rate function.

Example 10.1.4 (Using MINITAB and R) *Suppose that the lifetime T is modeled by the Weibull distribution whose threshold parameter is $\tau = 0$, shape parameter $\beta = 0.5$, and the scale parameter $\alpha = 40$. Find the reliability and hazard rate function at $T = t$, where $t = 100$.*

MINITAB

Using Minitab (see Chapter 5), we have



Probability Density Function

Weibull with shape = 0.5 and scale = 40

x	f(x)
100	0.0016265

Cumulative Distribution Function

Weibull with shape = 0.5 and scale = 40

x	P(X ≤ x)
100	0.794259

The reliability at 100 is given by $R(100) = 1 - F(100) = 1 - 0.794259 = 0.205741$

Further, the hazard rate function at $t = 100$ is given by

$$h(100) = f(100)/R(100) = 0.0016265/0.205741 = 0.0079$$

USING R

R has a built in Weibull density function 'dweibull(x, shape, scale)' and a Weibull c.d.f. 'pweibull(q, shape, scale)', where both 'x' and 'q' represent the quantile and 'shape' and 'scale' are the shape and scale parameters of the Weibull distribution, respectively. Referring to information provided in Example 10.1.4, we proceed as follows.

```
f.100 = dweibull(x = 100, shape = 0.5, scale = 40)
F.100 = pweibull(q = 100, shape = 0.5, scale = 40)


#The reliability at t = 100
R.100 = 1−F.100
R.100 #R output
[1] 0.2057407


#The hazard rate function at t = 100
h.100 = f.100/R.100
h.100 #R output
[1] 0.007905694
```

The *lognormal* is another distribution that is used to study reliability problems in engineering, medicine, and other fields. Recall that if the random variable T is distributed as *lognormal* with parameters μ and σ (i.e., $\ln T$ is distributed as normal with mean μ and standard deviation σ), then its p.d.f. is given by

$$f(t) = \begin{cases} \frac{1}{\sigma t\sqrt{2\pi}} e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}, & t > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{10.1.13}$$

From the definition of lognormal, it can easily be seen that the reliability function

$$R(t) = P(T \geq t) = 1 - P(T \leq t)$$

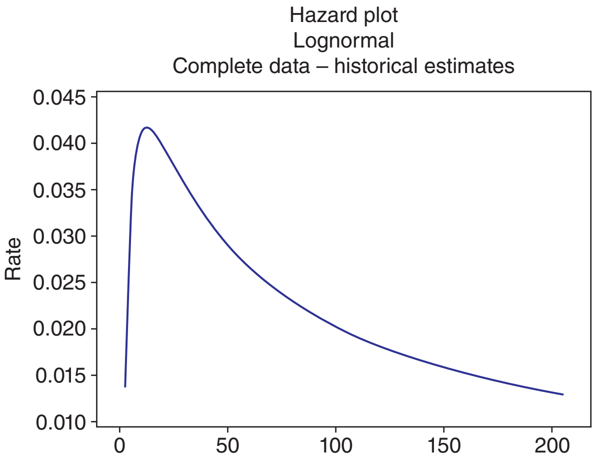


Figure 10.1.4 Hazard rate function for a lognormal distribution with parameters $\mu = 3$ and $\sigma = 1$.

for the lognormal,

$$R(t) = 1 - P(\ln T \leq \ln t) = 1 - P(Y \leq \ln t) \tag{10.1.14}$$

where Y is normally distributed with mean μ and variance σ^2.

The hazard rate function for the lognormal distribution is neither strictly increasing nor strictly decreasing, but first increases and then decreases. For example, the hazard rate function for a lognormal distribution with parameters $\mu = 3$ and $\sigma = 1$ is as shown in Figure 10.1.4.

Example 10.1.5 (Hazard rate for breast cancer patients) *Suppose that the lifetime in months of breast cancer patients after a mastectomy, followed by radiation and chemotherapy, is modeled by the lognormal distribution with parameters $\mu = 3$ and $\sigma = 1$. Find the reliability and the hazard rate function for these patients at five years (60 months).*

Solution: Using equation (10.1.13), we find that $(\mu = 3, \sigma = 1)$

$$f(60) = 0.0036535, F(60) = 0.863098$$

Now from equation (10.1.14), we have that

$$R(60) = 1 - F(60) = 1 - 0.863098 = 0.136902$$

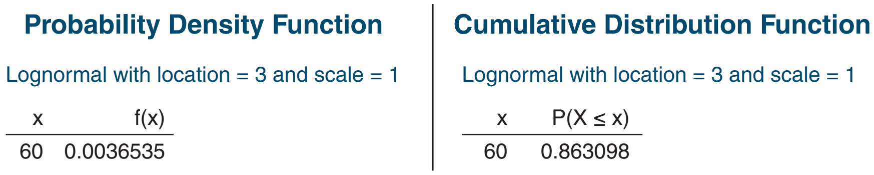
and hence,

$$h(60) = f(60)/R(60) = 0.0036535/0.136902 = 0.0267$$

To obtain this result using MINITAB and R, we proceed as follows:

MINITAB

Using Minitab (see Chapter 5), we have



Probability Density Function

Lognormal with location = 3 and scale = 1

x	f(x)
60	0.0036535

Cumulative Distribution Function

Lognormal with location = 3 and scale = 1

x	P(X ≤ x)
60	0.863098

The reliability at 60 is given by

$$R(60) = 1 - F(60) = 1 - 0.863098 = 0.136902$$

Further, the hazard rate function at $t = 60$ months is given by

$$h(60) = f(60)/R(60) = 0.0036535/0.136902 = 0.0267$$

USING R

R has a built in lognormal density ‘dlnorm(x, meanlog, sdlog)’ and a lognormal cumulative distribution function ‘plnorm(q, meanlog, sdlog)’, where both ‘x’ and ‘q’ represent the quantile, and ‘meanlog’ and ‘sdlog’ are the mean and the standard deviation of the lognormal distribution, respectively.

```
f.60 = dlnorm(x = 60, meanlog = 3, sdlog=1)
F.60 = plnorm(q = 60, meanlog = 3, sdlog=1)


#The reliability at t = 100
R.60 = 1-F.60
R.60 #R output
[1] 0.1369019


#The hazard rate function at t = 100
h.60 = f.60/R.60
h.60 #R output
[1] 0.02668676
```

The above results imply that only about 13.69% of the patients survive five years. The hazard rate function is given by (here Y is normally distributed with $\mu = 3$ and $\sigma = 1$)

$$h(60) = f(60)/R(60) = 0.0036535/0.1369 = 0.0267$$

That is the instantaneous death rate at $t = 60$ months is approximately 0.0267.

For another example, we consider a chain of n links, each link having strength X, where X is a random variable with exponential p.d.f. $f(x) = \lambda e^{-\lambda x}$. Clearly, the strength of the chain depends on the strength of its weakest link, and we are led to consider the distribution of the smallest chain strength in a sample of n links. In other words, the chain will break as soon as the weakest link breaks, or the time to failure for the chain is the same as for the weakest link. This situation also arises when several components in a system are placed in series so that the system fails as soon as the first component fails. The general expression for the c.d.f. G of the smallest strength in a sample of n (see equation (7.4.11)) is given by (please note that equation (7.4.11) is available on the book website: www.wiley.com/college/gupta/statistics2e)

$$G(x) = 1 - [1 - F(x)]^n \tag{10.1.15}$$

where $F(x)$ is the c.d.f. of the exponential. Since here

$$F(x) = 1 - e^{-\lambda x}$$

we obtain

$$G(x) = 1 - e^{-n\lambda x} \tag{10.1.16}$$

Differentiating this c.d.f. gives the p.d.f. for the smallest order statistic in a sample of n items from an exponential distribution, and we obtain

$$g(x) = n\lambda e^{-n\lambda x} \tag{10.1.17}$$

It can easily be verified that the reliability and hazard rate functions at $T = t$ are given by

$$R(t) = 1 - G(t) = e^{-n\lambda t} \tag{10.1.18}$$

$$h(t) = g(t)/R(t) = n\lambda \tag{10.1.19}$$

10.1.2 Employing the Hazard Function

The choice of an appropriate life distribution usually depends on what the engineering expert selects for the hazard rate of the items, $h(t)$. Knowing the approximate form of the hazard rate function, one can derive an appropriate form for the life distribution. To elucidate this fact, first recall that the hazard rate function h is defined as

$$h(u) = \frac{f(u)}{1 - F(u)} \quad \text{where} \quad \frac{d}{du}F(u) = f(u) \tag{10.1.20}$$

We then have

$$h(u)du = \frac{f(u)du}{1 - F(u)} = \frac{dF(u)}{1 - F(u)} \tag{10.1.21}$$

so that

$$\int_0^t h(u)du = \int_0^t \frac{1}{1 - F(u)}dF(u) = -\ln[1 - F(u)]\Big|_0^t \tag{10.1.22}$$

or

$$-\int_0^t h(u)du = \ln\frac{1 - F(t)}{1 - F(0)} = \ln[1 - F(t)] = \ln[R(t)] \tag{10.1.23}$$

since $F(0) = 0$. Furthermore, solving equation (10.1.23) for $R(t)$ we have,

$$R(t) = \exp\left[-\int_0^t h(u)du\right] \tag{10.1.24}$$

Now, from equation (10.1.20), we have $f(t) = h(t)[1 - F(t)] = h(t)R(t)$, so we can write, using equation (10.1.24), that

$$f(t) = h(t) \times \exp\left(-\int_0^t h(u)du\right) \tag{10.1.25}$$

or

$$f(t) = h(t) \times \exp(-H(t)) \tag{10.1.26}$$

where $H(t)$ is the cumulative hazard rate function given by $H(t) = \int_0^t h(u)du$.

We note that when $h(t) = \text{constant} = \lambda$, then $f(t) = \lambda e^{-\lambda t}$, the exponential distribution. Further, if $h(t)$ is a simple power function of t, say $h(t) = \eta t^{\eta-1}$, then $f(t) = \eta t^{\eta-1}\exp(-t^\eta)$, which is the Weibull distribution with $\beta = \eta, \alpha = 1$, and $\tau = 0$. If $h(t) = e^t$, then $f(t) = e^t\exp[-(e^t - 1)]$, a distribution first used by Gumbel (1958).

PRACTICE PROBLEMS FOR SECTION 10.1

1. Suppose that the lifetime in years of fuel pumps used in an aircraft gas turbine engine is modeled by the Weibull distribution, with threshold parameter 0, shape parameter $\beta = 1.0$, and scale parameter $\alpha = 30$. Find the reliability and the hazard rate function for these pumps at 10 years.
2. Suppose that the lifetime, in months, of heart patients after quadruple bypass surgery is modeled by the gamma distribution with shape parameter 3.5 and scale parameter 20. Find the reliability and the hazard rate function for these patients at six years (72 months).
3. The time T to failure of a computer hard drive is exponentially distributed with mean time to failure $\mu = 1/\lambda = 5000$ hours. Find the reliability and the hazard rate function for these hard drives at 3000 hours.
4. In Problem 2, suppose that the lifetime in months of heart patients after quadruple bypass surgery is modeled by the lognormal distribution with $\mu = 4$ and $\sigma = 1$. Find the reliability and the hazard rate function for these patients at six years (72 months).
5. Suppose that the hazard function of the transmission of a luxury car is given by

$$h(t) = \frac{\alpha\beta t^{\beta-1}}{1+\alpha t^\beta}, \quad \alpha > 0, \beta > 0, t \geq 0$$

   Find the density, reliability, and cumulative hazard function of the life of the transmission.
6. Suppose that in Problem 5, $\alpha = 1$ and $\beta = 0.5$. Find the reliability and the hazard rate at time $T = t$, where $t = 10$.

10.2 ESTIMATION: EXPONENTIAL DISTRIBUTION

The exponential distribution $f(t) = \lambda e^{-\lambda t}$ and its resultant constant hazard rate λ are very frequently employed in reliability engineering. We turn now to the problem of estimating λ, the hazard rate, from data gathered from a life test.

Suppose that a random sample of n items representative of standard production items are placed in usage and stressed in a fashion representative of the stresses encountered by the items when in common use. For example, n light bulbs may be turned on, and the time to failure of each bulb is noted. Often, instead of waiting until all n items have failed, the test is ended at the time of the kth failure, $k \leq n$, where k is chosen in advance. This method of testing is called *type II censoring* (type I censoring is discussed following Example 10.2.1). The times of the k failures are naturally ordered as we see them, and denoting the time to failure of the jth item, $1 \leq j \leq k$, by $t_{(j)}$, we have

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)} \tag{10.2.1}$$

Note that the only information we have about the other $(n-k)$ times to failure is that they exceed $t_{(k)}$. Now, using the methods of Chapter 7, it is easy to deduce that the joint density function of the first k-order statistics in a random sample of n, that is, $t_{(1)}, \ldots, t_{(k)}$, is given by

$$\frac{n!}{(1!)^k(n-k)!} f(t_{(1)}) \cdots f(t_{(k)})[1 - F(t_{(k)})]^{n-k} \tag{10.2.2}$$

where $F(x)$ is the c.d.f. corresponding to $f(x)$. In fact, given n, k, and the values of $t_{(j)}$, $1 \leq j \leq k$, equation (10.2.2) is the likelihood function (see Chapter 8) based on an experiment in which there is type II censoring as soon as the kth item fails. If we now substitute $f(t) = \lambda \exp(-\lambda t)$ and $F(t) = 1 - \exp(-\lambda t)$ in the joint p.d.f. given by equation (10.2.2) and invoke the principle of maximum likelihood, the maximum likelihood estimator (MLE) $\hat{\lambda}$ of λ is found to be

$$\hat{\lambda} = \frac{k}{T_k}, \quad T_k = \sum_{i=1}^{k} t_{(i)} + (n-k)t_{(k)} \tag{10.2.3}$$

Here, T_k is an estimate of the total of the failure times of all n items, based on the available k times. Now, remembering that the mean of the exponential distribution used here is $\mu = 1/\lambda$, we find that the MLE of the mean of the time between failures is

$$\hat{\mu} = \frac{T_k}{k} \tag{10.2.4}$$

To obtain a confidence interval for μ, it has been shown (see Halperin, 1952, or Mann et al., 1974) that the quantity $2T_k/\mu$ is a random variable distributed as a chi-square variable with $v = 2k$ degrees of freedom. We then have that

$$P\left(\chi^2_{2k,1-a/2} \leq \frac{2T_k}{\mu} \leq \chi^2_{2k,\alpha/2}\right) = 1 - \alpha \tag{10.2.5}$$

Since $T_k = k\hat{\mu}$ from equation (10.2.4), we easily find that

$$\left[\frac{2k\hat{\mu}}{\chi^2_{2k,\alpha/2}}, \frac{2k\hat{\mu}}{\chi^2_{2k,1-\alpha/2}}\right] \tag{10.2.6}$$

is a $100(1-\alpha)\%$ confidence interval for μ. The corresponding $100(1-\alpha)\%$ confidence limits for the hazard rate (λ) are obtained by simply substituting $\lambda = 1/\mu$ and $\hat{\lambda} = 1/\hat{\mu}$ to give the lower and upper limits λ_L and λ_U as

$$\begin{cases} \lambda_L &= \frac{\hat{\lambda}}{2k}\chi^2_{2k,1-\alpha/2} \\ \lambda_U &= \frac{\hat{\lambda}}{2k}\chi^2_{2k,\alpha/2} \end{cases} \tag{10.2.7}$$

Now for the exponential distribution, the reliability is given by $R(t) = \exp(-\lambda t)$, so the MLE of the reliability function $\hat{R}(t)$ is given by

$$\hat{R}(t) = \exp(-\hat{\lambda}t) \tag{10.2.8}$$

Then, we use equation (10.2.7) and write the $100(1-\alpha)\%$ confidence interval for the reliability at any time t as

$$[\exp(-\lambda_U t), \exp(-\lambda_L t)] \tag{10.2.9}$$

where λ_L and λ_U are the lower and upper limits of λ given by equation (10.2.7).

Occasionally, one will read in the reliability literature that μ, the mean time between failures (MTBF), is estimated "with $100(1-\alpha)\%$ confidence." What is computed in these cases is the lower bound of the one-sided confidence interval for the parameter μ, that is the single quantity

$$2k\hat{\mu}/\chi^2_{2k,\alpha} \tag{10.2.10}$$

Clearly, merely by altering α, alternative values for the confidence estimate are possible, and we note that the higher the confidence coefficient $(1-\alpha)$, the smaller the proffered value for the MTBF.

Example 10.2.1 (Confidence interval for mean time between failures) *A random sample of 10 light bulbs is placed on life test, and the test concluded after the fourth failure. The recorded times to failure are 836, 974, 1108, and 1236 hours. Given that the time to failure of the bulbs is appropriately represented by an exponential distribution, the mean time to failure is estimated to be (see equations (10.2.3) and (10.2.4))*

$$\hat{\mu} = \frac{[836 + 974 + 1108 + 1236]}{4} + \frac{6(1236)}{4} = 2892.5 \text{ hours}$$

The estimated hazard rate is:

$$\hat{\lambda} = (2892.5)^{-1} = 3.4572 \times 10^{-4} \text{ failures per hour}$$

Further, from equation (10.2.6), 95% confidence limits for the MTBF μ are

$$\left(\frac{2(4)(2892.5)}{17.5346}, \frac{2(4)(2892.5)}{2.1797}\right) = (1319.67, 10616.14)$$

The 95% confidence limits for the hazard rate λ are

$$(9.42 \times 10^{-5}, 75.78 \times 10^{-5})$$

From equation (10.2.10), the mean time to failure with 95% confidence is given by $2(4)(2892.5)/15.5073 = 1492.2$ hours; that is the MTBF at 95% confidence is 1492.2 hours (this is the lower bound of a one-sided confidence interval with confidence coefficient 95%). Observe now that the estimated mean time to failure with 99% confidence is less than 1492.2, and the 99% confidence estimate is $2(4)(2892.5)/20.0902 = 1151.81$ hours.

From these data, the estimated reliability of the bulbs at $t = 2000$ hours is $\hat{R} = e^{-t/\hat{\mu}} = e^{-2000/2892.5} = 0.50$.

That is, it is estimated that 50% of the bulbs will remain workable at 2000 hours. The estimated reliability at $t = 2000$ hours, at 95% confidence is given by $\exp(-t/\hat{\mu}) = \exp(-2000/1492.2) = 0.26$. Thus, the reliability engineer is 95% confident that 26% of the bulbs will be operating satisfactorily at 2000 hours.

The maximum likelihood estimate of the time at which the reliability attains the value $R = 0.75$, found solving $R(t) = e^{-\lambda t} = e^{-t/\mu}$ for t, say $\hat{t}$, given by

$$\hat{t} = \hat{\mu}\ \ln\left(\frac{1}{R}\right) = (2892.5)\ln\left(\frac{1}{0.75}\right) = 832.1 \text{ hours}$$

The corresponding estimate of the time, for $R = 0.75$ at 99% confidence, is

$$\hat{t} = (1151.8)\ln\left(\frac{1}{0.75}\right) = 331.4 \text{ hours}$$

An alternative life testing scheme, termed *type I censoring*, is to stop the test of the sample of n items at some fixed time t_0 decided on in advance. In these circumstances, the estimate of the time to failure is

$$\hat{\mu} = \frac{1}{k}\left(\sum_{i=1}^{k} t_{(i)} + (n-k)t_0\right) = \frac{T_0}{k}, \qquad k > 0 \tag{10.2.11}$$

where $t_{(1)}, \ldots, t_{(k)}$ are the observed failure times that are less than t_0. Approximate $100(1-\alpha)\%$ confidence limits for μ are obtained using equation (10.2.5).

Example 10.2.2 (Example 10.2.1 continued) *Referring to Example 10.2.1, suppose that it had been decided to complete the life test after 1250 hours and that only the four failures recorded in Example 10.2.1 had occurred in that time, all $n - k = 6$ others having lifetime greater than 1250 hours. (this sometimes called "right censoring").*

Then,

$$\hat{\mu} = \frac{[836 + 974 + 1108 + 1236]}{4} + \frac{6(1250)}{4} = 2913.5 \text{ hours}$$

The approximate confidence limits for the MTBF μ are 1329.3 and 10693.2. (These are found using equation (10.2.6) with $\hat{\mu}$ as above.)

In addition to the exponential distribution, the Weibull and lognormal distributions are frequently used in reliability theory. In Example 10.2.3, we employ all three distributions. Then, in Section 10.4, we resume discussion of the use of the Weibull distribution.

Example 10.2.3 (Using MINITAB and R) *Referring to the data in Example 10.2.2 and using MINITAB and R, apply the exponential, Weibull, and lognormal models to the data.*

MINITAB

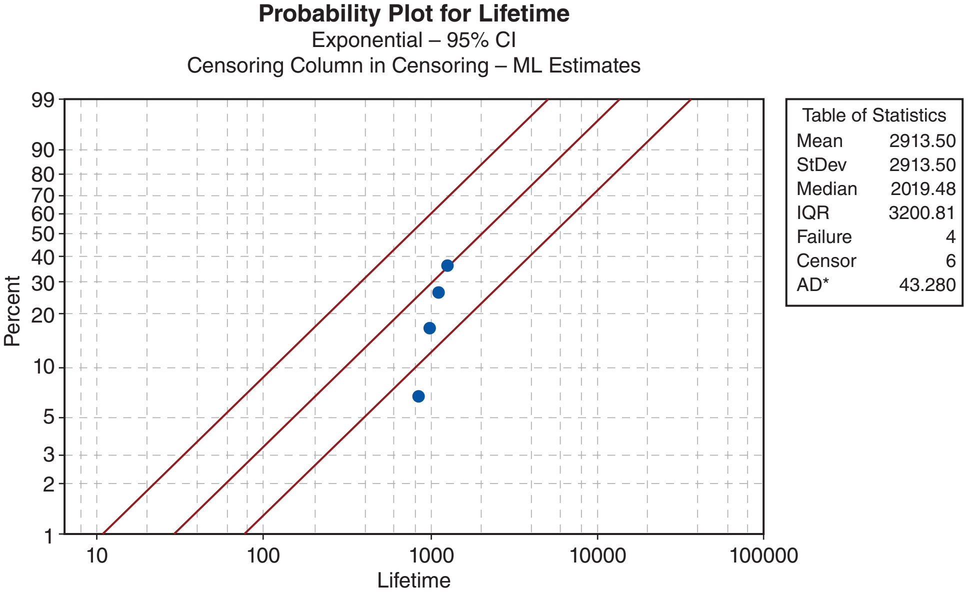
To fit exponential, Weibull, and lognormal distributions we proceed as follows:

1. Enter the data in column C1 and create a censoring column C2 by entering 1 for uncensored observations and 0 for censoring observations. (If an observation in survival data does not possess a characteristic of interest then it is called a censored observation. For example, a machine is still functioning or a patient is still alive at the end of the study period.)
2. From the Menu bar, select **Stat** > **Reliability/Survival** > **Distribution Analysis (Right Censoring)** > **Parametric Distribution Analysis**.
3. A dialog box **Parametric Distribution Analysis-Right Censoring** appears. In this dialog box, enter the data column under variables. Then, select the **Censor** option and enter the censoring column in the box under **Use censoring columns** and click **OK**. Now select other options, for example, **Estimate...**, and select either estimation method; least-squares or the maximum likelihood method. Then, under **Graphs...** select any desired plots such as Probability plot, Survival plot, Cumulative Failure plot, Display confidence intervals on above plots, Hazard plot and then click **OK**.
4. Finally, select one of the **Assumed Distributions** by clicking on the pull-down arrow.
5. Click **OK**. The final results appear in the Session window as shown below.

1. **Exponential Model**

```
Parameter Estimates

                      Standard    95.0% Normal CI
Parameter  Estimate   Error       Lower     Upper
Mean       2913.50    1456.75     1093.49   7762.75


Log-Likelihood = –35.908

Goodness-of-Fit

Anderson-Darling
      (Adjusted)
          43.280
```

```
Characteristics of Distribution

                                      Standard    95.0% Normal CI
                           Estimate   Error       Lower     Upper
Mean(MTTF)                 2913.50    1456.75     1093.49   7762.75
Standard Deviation         2913.50    1456.75     1093.49   7762.75
Median                     2019.48    1009.74     757.949   5380.73
First Quartile(Q1)         838.162    419.081     314.577   2233.20
Third Quartile(Q3)         4038.97    2019.48     1515.90   10761.5
Interquartile Range(IQR)   3200.81    1600.40     1201.32   8528.26
```



```
Table of Statistics
Mean      2913.50
StDev     2913.50
Median    2019.48
IQR       3200.81
Failure         4
Censor          6
AD*        43.280
```

Note that the probability plots are used to assess whether a particular distribution fits the data. In general, the closer the points fall to the fitted line, the better the fit. The points are plotted using median rank (called Bernard's approximation), that is $[(i - 0.3/n + 0.4),\ t_{(i)}]$, $i = 1, 2, 3, 4$.

2. **Weibull Model**

Parameter Estimates

		Standard	95.0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Shape	5.74677	2.68328	2.30134	14.3505
Scale	1403.25	151.747	1135.24	1734.54

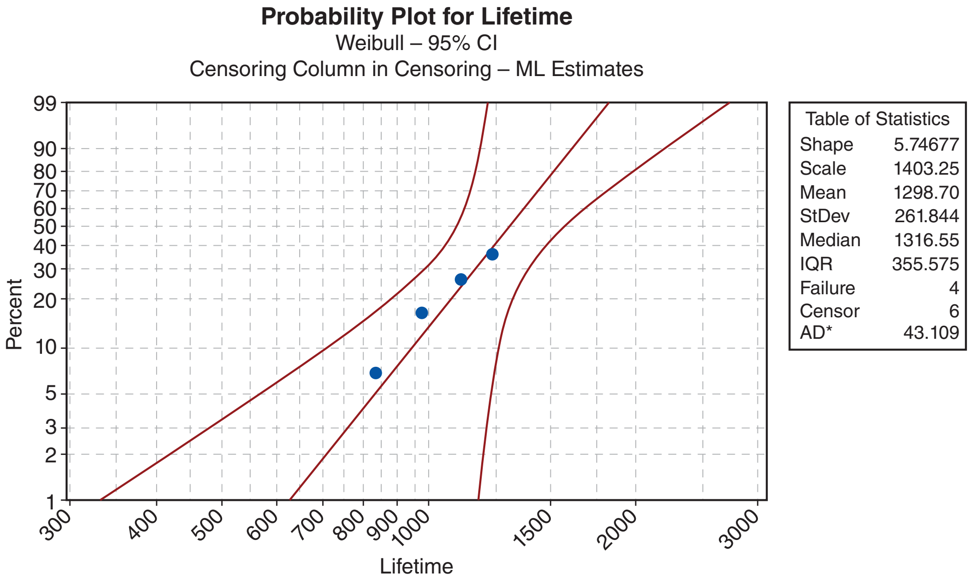
Log-Likelihood = –31.907

Goodness-of-Fit

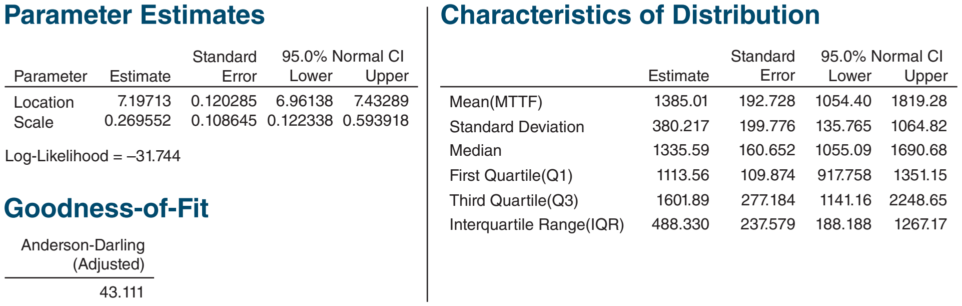
Anderson-Darling (Adjusted)
43.109

Characteristics of Distribution

		Standard	95.0% Normal CI	
	Estimate	Error	Lower	Upper
Mean(MTTF)	1298.70	123.310	1078.17	1564.33
Standard Deviation	261.844	124.482	103.128	664.827
Median	1316.55	123.195	1095.94	1581.57
First Quartile(Q1)	1129.75	106.816	938.643	1359.76
Third Quartile(Q3)	1485.32	186.744	1160.92	1900.37
Interquartile Range(IQR)	355.575	179.305	132.342	955.355



3. Lognormal Model



Parameter Estimates

Parameter	Estimate	Standard Error	95.0% Normal CI Lower	95.0% Normal CI Upper
Location	7.19713	0.120285	6.96138	7.43289
Scale	0.269552	0.108645	0.122338	0.593918

Log-Likelihood = –31.744

Goodness-of-Fit

Anderson-Darling (Adjusted)
43.111

Characteristics of Distribution

	Estimate	Standard Error	95.0% Normal CI Lower	95.0% Normal CI Upper
Mean(MTTF)	1385.01	192.728	1054.40	1819.28
Standard Deviation	380.217	199.776	135.765	1064.82
Median	1335.59	160.652	1055.09	1690.68
First Quartile(Q1)	1113.56	109.874	917.758	1351.15
Third Quartile(Q3)	1601.89	277.184	1141.16	2248.65
Interquartile Range(IQR)	488.330	237.579	188.188	1267.17

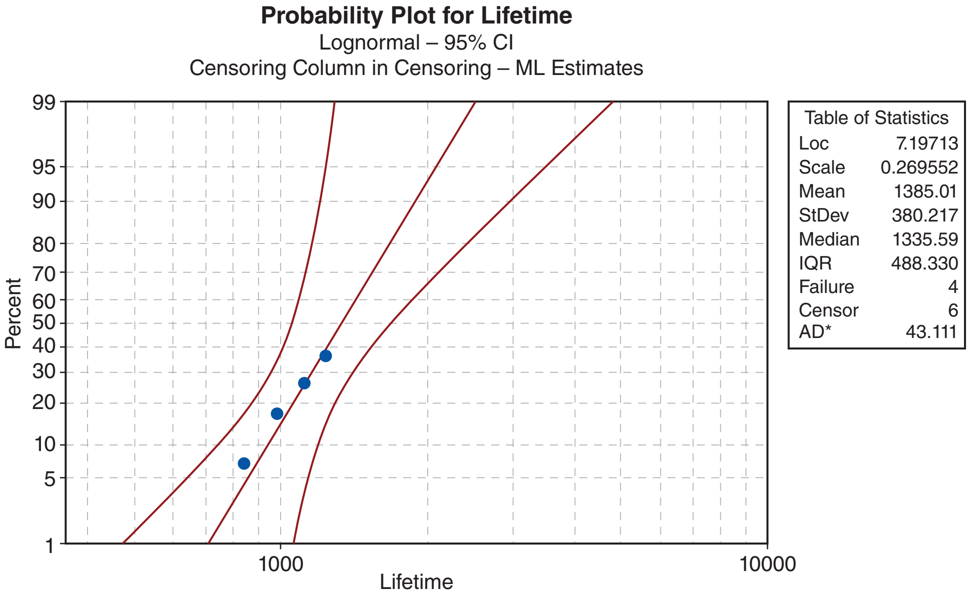


Table of Statistics	
Loc	7.19713
Scale	0.269552
Mean	1385.01
StDev	380.217
Median	1335.59
IQR	488.330
Failure	4
Censor	6
AD*	43.111

USING R

To fit survival models in R, we may use ‘survreg()’ function in library ‘survival’. However, it is required to change the data to a survival R object using the ‘Surv()’ function so that we can apply it in the ‘survreg()’ function as shown in the following R code. Note that the ‘survreg()’ function transforms the time variable to natural log (ln) scale and fit Accelerated Failure Time (AFT) models.

```
#Load survival package
library(survival)
#Failure times including six right censored observations
Time = c(836,974,1108,1236,1250,1250,1250,1250,1250,1250)
#Identify whether the observation is right censored (=0) or not (=1)
Status = c(1,1,1,1,0,0,0,0,0,0)
```

```
#Fit data with an Exponential distribution
model1 = survreg(Surv(Time,Status)~1, dist = “exponential”)
summary(model1)

#R summary output
```

	Value	Std. Error	z	p
(Intercept)	7.98	0.50	15.9	$<$ 2e-16

```
Exponential distribution

Loglik(model)= -35.9, Loglik(intercept only)= -35.9

#Fitted mean can be obtained via 1/exp(-Intercept of the model)
mean = 1/exp(-model1$icoef)

#Fit data with a Weibull distribution
model2 = survreg(Surv(Time,Status)~1, dist = “weibull”)
summary(model2)

#R summary output
```

	Value	Std. Error	z	p
(Intercept)	7.247	0.108	67.01	$<$ 2e-16
Log(scale)	−1.749	0.467	−3.75	0.00018

```
Weibull distribution

Loglik(model)= -31.9, Loglik(intercept only)= -31.9

#Fitted Shape can be obtained via 1/Scale
Shape = 1/model2$scale

#Fitted Scale can be obtained via 1/exp(-Intercept of the model)
Scale = 1/exp(-model2$icoef[1])

#Fit data with a Lognormal distribution
model3= survreg(Surv(Time,Status)~1, dist = “lognormal”)
summary(model3)

#R summary output
```

	Value	Std. Error	z	p
(Intercept)	7.197	0.120	59.83	$<$ 2e-16
Log(scale)	−1.311	0.403	−3.25	0.0011

```
Log Normal distribution

Loglik(model)= -31.7, Loglik(intercept only)= -31.7


#Fitted Location can be obtained via intercept of the model
Location = model3$icoef[1]


#Fitted Scale can be obtained via 1/exp(-Log(scale) of the model)
Scale = 1/exp(-model3$icoef[2])
```

To see which model fits the given data better, we use as criterion the Anderson–Darling statistic (see Anderson and Darling, 1952) in MINITAB and log-likelihood ratio test statistic (see Lee and Wang, 2003, p. 233) in R. Comparing the values of both the Anderson–Darling statistic and the log-likelihood ratio test, we find that Weibull fits negligibly better than lognormal and exponential while the lognormal fits better than the exponential.

Note that the smallest value of the Anderson–Darling statistic is an indication that the corresponding model gives rise to the best fit of the data among the three possible models. Similarly, the smallest absolute value of the log-likelihood ratio test conveys the same information. Thus, in this case, even using the log-likelihood ratio test, we can conclude that both Weibull and lognormal distributions are preferred to the exponential distribution.

PRACTICE PROBLEMS FOR SECTION 10.2

1. A random sample of 12 mechanical parts of a system is placed on life test, and the test is concluded after the sixth failure. The recorded times to failure are 964, 1002, 1067, 1099, 1168, and 1260 hours. Assuming that the time to failure of the parts is appropriately represented by an exponential distribution, determine the MLE estimate of the mean time to failure.
2. Referring to Problem 1, estimate the hazard rate and find a 95% confidence interval for the hazard rate.
3. Referring to Problem 1, construct 95% confidence interval for the mean time to failure. Also find a 95% confidence interval for the reliability of the part at $t = 1180$ hours.
4. A random sample of 10 automobile voltage regulators is placed on life test, and it is decided to complete the life test after 78 months. The recorded times to failure are 58, 62, 68, 74, and 77 months. Assuming that the time to failure of the voltage

regulators is appropriately represented by an exponential distribution, estimate the mean time to failure.

5. Referring to Problem 4, estimate the hazard rate and find a 99% confidence interval for the hazard rate.
6. Referring to Problem 4, find a 99% confidence interval for the mean time to failure. Also find a 99% confidence interval for the reliability of the part at $t = 70$ months.

10.3 HYPOTHESIS TESTING: EXPONENTIAL DISTRIBUTION

If we wish to test a hypothesis about μ, then we know that the mean time between failure of items has an *exponential life distribution*, so that one technique is to use the data from a life test of n items and construct a $100(1-\alpha)\%$ confidence interval for μ (one-sided or two-sided). If the hypothesized value μ_0 lies outside the interval, the hypothesis may be rejected with a type I error having probability equal to α. This approach, however, is not useful when one must plan an experimental trial to test, with appropriate α and β risks, the hypothesis:

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu = \mu_1 < \mu_0$$

Interestingly, this problem may be stated in terms of $R(t_0)$, the reliability of the items at some fixed point in time t_0 called the "mission time." For example, a typical scenario declares that items are acceptable if their reliability is 0.90 at $t_0 = 1000$ hours, but the item should be rejected if their reliability is 0.75 at $t_0 = 1000$ hours. Then, we may state the hypothesis corresponding to H_0 above as

$$H_0 : R_0(t_0) = R_0(1000) = 0.9 \quad \text{versus} \quad H_1 : R_1(t_0) = R_1(1000) = 0.75$$

where we note that $R_1(t_0) < R_0(t_0)$.

But we are dealing with the case that lifetimes of the items are exponentially distributed, so $R(t) = e^{-t/\mu}$. Hence, the corresponding values for μ are $\mu_0 = -t_0/\ln[R(t_0] = -1000/\ln(0.9) = 9491$ and $\mu_1 = -1000/\ln(0.75) = 3476$. For convenience, we use $\mu_0 = 9500$ and $\mu_1 = 3500$ hours. We are then in the situation that we wish to test

$$H_0 : \mu = \mu_0 = 9500 \quad \text{versus} \quad H_1 : \mu = \mu_1 = 3500$$

with the probability of type I error $= \alpha$ and the probability of type II error $= \beta$.

As discussed in Section 9.12, there is a sequential test of this hypothesis. To conduct such a test, we place n items on test and record the time to failure t of each item. Let T_m equal the total accumulated time after m items fail; that is $\sum_{i=1}^{m} t_i$. Given that the failure times t_i are independent events from an exponential distribution $f(t) = \mu^{-1}\ \exp(-t/\mu)$, the sequential test may be stated as follows:

Continue sampling as long as T_m is such that

$$A \approx \frac{\beta}{1-\alpha} \leq W = \frac{(\mu_1)^{-m} e^{-T_m/\mu_1}}{(\mu_0)^{-m} e^{-T_m/\mu_0}} \leq \frac{1-\beta}{\alpha} \approx B \tag{10.3.1a}$$

But if $W < A$ for some m, stop sampling and *do not reject*

$$H_0 \ :\ \mu = \mu_0 \tag{10.3.1b}$$

Finally, if W >B for some m, stop sampling and *reject*

$$H_0 \ :\ \mu = \mu_0 \quad \text{in favor of} \quad H_1 \ :\ \mu = \mu_1 \tag{10.3.1c}$$

We remind the reader that $\mu_1 < \mu_0$.

This comes about naturally when we see that W is the ratio of the likelihood of μ_1 to the likelihood of μ_0 based on the observed lifetimes. If W is "too large," for example, then the likelihood of μ_0 dominates the likelihood of μ_1, an indication that we should reject $H_0 :\ \mu = \mu_0$ in favor of $H_1 : \mu = \mu_1$, and so on, "too large" and "too small" are defined, respectively, by $W > B = (1-\beta)/\alpha$ and $W < A = \beta/(1-\alpha)$, which gives the above test procedure the property of the probability of type I error $= \alpha$ and the probability of type II error $= \beta$ (power of the test $= 1-\beta$). We note that $A < B$ for $\alpha < 0.5$ and $\beta < 0.5$ (the usual case).

Now after some algebra, we can restate equations (10.3.1a)–(10.3.1c) as:

Continue sampling as long as T_m lies between the lines

$$a_0 + bm \quad \text{and} \quad a_1 + bm$$

where

$$a_0 = \frac{\mu_1\mu_0}{\mu_1 - \mu_0}\ln(B), \quad a_1 = \frac{\mu_1\mu_0}{\mu_1 - \mu_0}\ln(A), \quad b = \frac{\mu_1\mu_0}{\mu_1 - \mu_0}\ln\left(\frac{\mu_1}{\mu_0}\right) \tag{10.3.1d}$$

with $\mu_1 < \mu_0$.

But if $T_m < a_0 + bm$, then reject $H_0\ : \mu = \mu_0$ in favor of $H_1\ : \mu = \mu_1$, but if $T_m > a_1 + bm$, then do not reject $H_0\ : \mu = \mu_0$. We illustrate this with the following example.

Example 10.3.1 (Sequential test of hypothesis) *Suppose for the situation in this section that we are dealing with the case* $\mu_0 = 9500, \mu_1 = 3500, \alpha = 0.05$ *and* $\beta = 0.10$.

Solution: Referring to equations (10.3.1a)–(10.3.1d), we find that

$$a_0 = -16{,}017, \quad a_1 = 12{,}476, \quad b = 5534$$

Thus, when plotting T_m versus m, as long as T_m falls between the two parallel lines $a_j + bm,\ j = 0, 1$, that is if T_m is such that

$$-16{,}017 + 5534m \leq T_m \leq 12{,}476 + 5534m$$

then sampling continues, but if T_m fall below $-16{,}017 + 5534m$, we reject H_0 in favor of $H_1\ : \mu = \mu_1 = 3500$, while if T_m falls above $12{,}476 + 5534m$, we do not reject H_0: $\mu_0 = 9500$.

In setting up this sequential decision plan, we assume that n, the number of items on test, is very large, or that as each item fails, it is immediately replaced on test by a new item. When the items are not replaced on failure, the same equations hold as above, except that at any time t we have

$$T_m = \sum_{i=1}^{m} t_{(i)} + (n - m)(t - t_m) \tag{10.3.2}$$

The sequential test plan above assumes that the individual times to failure will be known. More commonly, n items are placed on a life test, and at some designated time t_p, the number of failed items x is recorded. Based on this evidence, a decision to accept or reject H_0 is taken.

10.4 ESTIMATION: WEIBULL DISTRIBUTION

We have noted that when the hazard function $h(t) = f(t)/[1 - F(t)]$ is a constant λ, the distribution of time to survival is exponential. However, for many manufactured items $h(t)$ may either rise or fall over an extended period of time. If we employ a simple power function $h(t) = \beta t^{\beta-1}$, then for $\beta > 0$, the hazard function will rise, but for $\beta < 0$, the hazard will decline as time increases. And since

$$f(t) = h(t) \exp\left[-\int_0^t h(x)dx\right] \tag{10.4.1}$$

the corresponding distribution of time to failure is the *standard Weibull distribution*

$$f(t|\beta) = \beta t^{\beta-1} e^{-t^{\beta}}$$

which is a member of the class of distributions given by

$$f(t|\alpha,\beta,\tau) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{t-\tau}{\alpha}\right)^{\beta-1} e^{-[(t-\tau)/\alpha]^{\beta}}, & \text{for } t > \tau \\ 0, & \text{otherwise} \end{cases} \tag{10.4.2}$$

where $\alpha > 0, \beta > 0,$ and $\tau \geq 0$.

We say that β is the *shape parameter*, α the *scale parameter*, and τ the *threshold parameter*.

If instead we take $y = [(t - \tau)/\alpha]^{\beta}$, the Weibull variable is transformed to an exponential variable whose p.d.f. is $f(y)dy = \lambda e^{-\lambda y}\, dy$. The reliability function for the Weibull is $R(t) = e^{-[(t-\tau)/\alpha]^{\beta}}$. Setting the threshold parameter at $\tau = 0$, and taking logarithms twice give $\ln\ln[1/R(t)] = \beta \ln t - \beta \ln \alpha$, which we rearrange to read

$$\ln t = \ln \alpha + (1/\beta)[\ln\ln(1/R(t))]$$

This is the equation of a straight line in the variables $(\ln\ln(1/R(t)),\ \ln t)$ whose intercept is $\ln(\alpha)$ and whose slope is $1/\beta$.

Suppose now that n items are placed on life test and that the observed failure times of the first k are recorded, $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$. As earlier, when order statistics were used to provide probability plots (see Chapter 5), an estimate of the proportion of working items available up to time $t_{(i)}$ is given by $\hat{F}(t_{(i)}) = (i - 1/2)/n$. The estimated reliability at time $t_{(i)}$ is then $\hat{R}(t_{(i)}) = 1 - \hat{F}(t_{(i)})$. Other schemes for estimating $\hat{F}(t_{(i)})$ are in common use, for example, setting $F(t_{(i)}) = i/(n+1)$, the so-called *mean rank*, or setting $F(t_{(i)}) = (i - 0.3)/(n + 0.4)$, called the *median rank*. Thus, if the values of $\ln t_{(i)}$ are plotted along the ordinate axis versus $\ln\ln(1/\hat{R})$ along the abscissa, a series of points lying reasonably well along a straight line is obtained when the times to failure are Weibull distributed.

However, if the axes are reversed, the slope estimates β directly, which we now denote by $\hat{\beta}$. Since $E(T) = \mu = \alpha \times \Gamma(1 + 1/\beta)$, the estimator of α is obtained using $\hat{\beta}$ and the average $\bar{t}$ to give $\hat{\alpha} = \bar{t}/\Gamma(1 + 1/\hat{\beta})$. The reader is referred to Mann et al. (1974), Lawless (2003), and Lee and Wang (2003) for further methods of estimation and tests of significance for the Weibull distribution.

Example 10.4.1 (Microchips) *A sample of* $n = 20$ *microchips are placed on life test, and the following* $k = 10$ *times to failure recorded:*

i	$t_{(i)}$	$\hat{F}(t) = (i - 0.5)/20$	$\hat{R}(t) = 1 - \hat{F}(t)$	ln $\ln(1/\hat{R}(t))$	$\ln t_{(j)}$
1	8.4	0.025	0.975	3.6762	2.1282
2	17.0	0.075	0.925	−2.5515	2.8332
3	19.1	0.125	0.875	−2.0134	2.9497
4	25.0	0.175	0.825	−1.6483	3.2189
5	29.2	0.225	0.775	−1.3669	3.3742
6	36.2	0.275	0.725	−1.1345	3.5891
7	43.8	0.325	0.675	−0.9338	3.7796
8	44.7	0.375	0.625	−0.7550	3.8000
9	65.4	0.425	0.575	−0.5917	4.1805
10	69.4	0.475	0.525	−0.4395	4.2399

Given that the Weibull distribution is appropriate to these data, a simple plot of the points $(\ln\ln(1/R(t)),\ \ln t)$ *on ordinary graph paper should provide a reasonably straight line, as in Figure 10.4.1. The fitted Weibull distribution is*

$$f(t) = \frac{1.45}{39.49}\left(\frac{t}{39.49}\right)^{0.45} \exp\left(-\left(\frac{t}{39.49}\right)^{1.45}\right)$$

Example 10.4.2 (Microchips) *Use the data of Example 10.4.1 and assume that the experimenter decided to terminate the test as soon as the 10th microchip failed, so that the time to failure for the rest of the chips is* 69.4+ *(right-censored data points). Using MINITAB, fit the Weibull model estimating the parameters by the maximum likelihood method and the least-squares method (discussed in detail in Chapter 15).*

MINITAB

Solution: Using the same steps as described in Example 10.2.3, we have

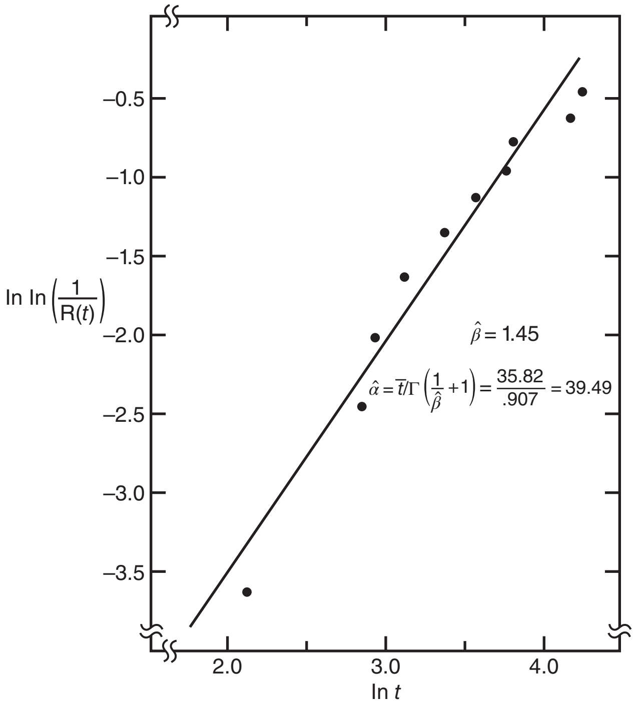


Figure 10.4.1 Graphical estimation of Weibull parameters.

Maximum Likelihood Method

Parameter Estimates

		Standard	95.0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Shape	1.41807	0.410116	0.804496	2.49960
Scale	89.0800	22.0897	54.7902	144.830

Log-Likelihood = –55.918

Goodness-of-Fit

Anderson-Darling (Adjusted)
70.717

Characteristics of Distribution

		Standard	95.0% Normal CI	
	Estimate	Error	Lower	Upper
Mean(MTTF)	81.0299	21.8602	47.7540	137.493
Standard Deviation	57.9508	27.5762	22.8040	147.268
Median	68.7912	15.5147	44.2139	107.030
First Quartile(Q1)	37.0010	9.85492	21.9535	62.3626
Third Quartile(Q3)	112.154	31.7971	64.3409	195.498
Interquartile Range(IQR)	75.1529	30.1175	34.2631	164.841

Least-Squares Method

Parameter Estimates

Parameter	Estimate
Shape	1.42424
Scale	83.7490

Log-Likelihood = –55.956

Goodness-of-Fit

Anderson-Darling (Adjusted)	Correlation Coefficient
70.737	0.986

Characteristics of Distribution

	Estimate
Mean(MTTF)	76.1316
Standard Deviation	54.2287
Median	64.7468
First Quartile(Q1)	34.9193
Third Quartile(Q3)	105.337
Interquartile Range(IQR)	70.4177

Observe that the value of the Anderson–Darling statistics is slightly smaller when we use ML estimates (see Figures 10.4.2 and 10.4.3).

If we use only the uncensored data to fit the Weibull model, then the resulting estimates, as shown below, are significantly different.

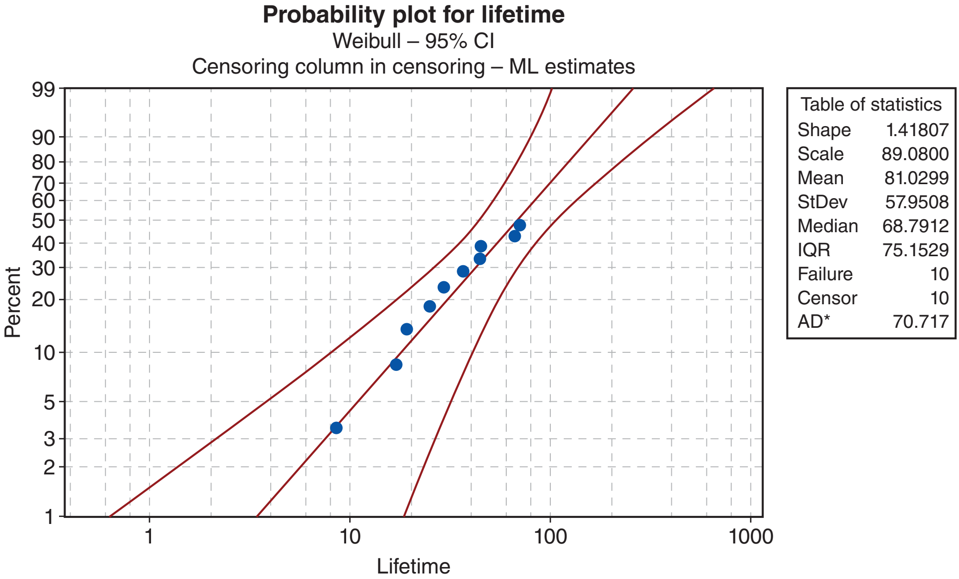


Table of statistics	
Shape	1.41807
Scale	89.0800
Mean	81.0299
StDev	57.9508
Median	68.7912
IQR	75.1529
Failure	10
Censor	10
AD*	70.717

Figure 10.4.2 Maximum Likelihood estimates and 95% CI for lifetime using the Weibull model.

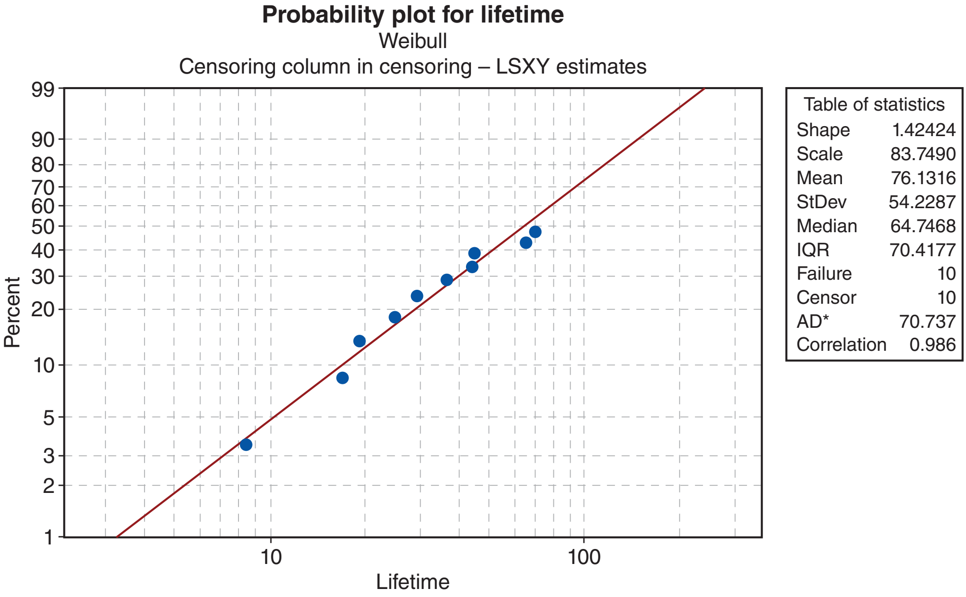
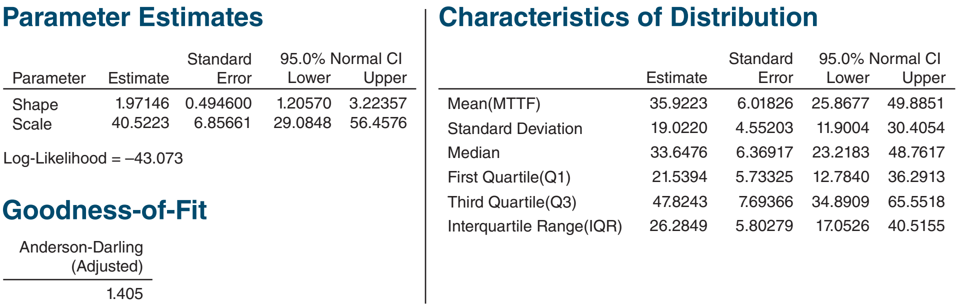


Table of statistics	
Shape	1.42424
Scale	83.7490
Mean	76.1316
StDev	54.2287
Median	64.7468
IQR	70.4177
Failure	10
Censor	10
AD*	70.737
Correlation	0.986

Figure 10.4.3 Least-squares estimates and 95% CI for lifetime using the Weibull model.

Maximum Likelihood Method (uncensored data)



Parameter Estimates

Parameter	Estimate	Standard Error	95.0% Normal CI Lower	Upper
Shape	1.97146	0.494600	1.20570	3.22357
Scale	40.5223	6.85661	29.0848	56.4576

Log-Likelihood = –43.073

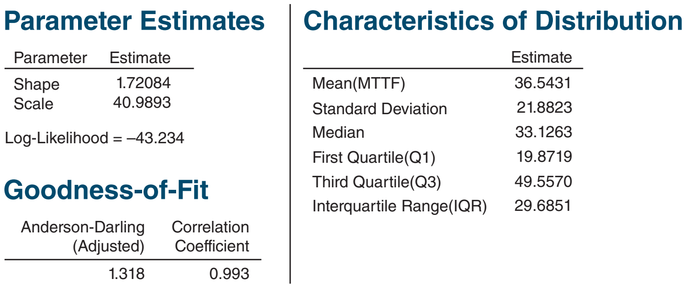
Goodness-of-Fit

Anderson-Darling (Adjusted)
1.405

Characteristics of Distribution

	Estimate	Standard Error	95.0% Normal CI Lower	Upper
Mean(MTTF)	35.9223	6.01826	25.8677	49.8851
Standard Deviation	19.0220	4.55203	11.9004	30.4054
Median	33.6476	6.36917	23.2183	48.7617
First Quartile(Q1)	21.5394	5.73325	12.7840	36.2913
Third Quartile(Q3)	47.8243	7.69366	34.8909	65.5518
Interquartile Range(IQR)	26.2849	5.80279	17.0526	40.5155

Least-Squares Method (uncensored data)



Parameter Estimates

Parameter	Estimate
Shape	1.72084
Scale	40.9893

Log-Likelihood = –43.234

Goodness-of-Fit

Anderson-Darling (Adjusted)	Correlation Coefficient
1.318	0.993

Characteristics of Distribution

	Estimate
Mean(MTTF)	36.5431
Standard Deviation	21.8823
Median	33.1263
First Quartile(Q1)	19.8719
Third Quartile(Q3)	49.5570
Interquartile Range(IQR)	29.6851

Estimation of Mean and Variance When the Time to Failure is Lognormal and the Data is Uncensored.

Suppose that the time to failure is lognormal with parameters μ and σ^2. Assume that a random sample of n parts is placed on a life test and the times to failure are t_i, $i = 1, 2, \ldots, n$. Then, it can be easily seen (see Chapter 8) that the MLEs of μ and σ^2 are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln\ t_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} (\ln\ t_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} \ln\ t_i\right)^2\right) \tag{10.4.3}$$

An unbiased estimator of σ^2 is given by $S^2 = (n/(n-1))\hat{\sigma}^2$ (see Chapter 8). Further it can be seen that $100(1-\alpha)\%$ confidence intervals for μ and σ^2 are given by

$$[\hat{\mu} - t_{(n-1),\ \alpha/2}(S/\sqrt{(n-1)}),\ \hat{\mu} + t_{(n-1),\ \alpha/2}(S/\sqrt{(n-1)})] \tag{10.4.4}$$

and

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{(n-1),\ \alpha/2}}, \frac{n\hat{\sigma}^2}{\chi^2_{(n-1),\ 1-\alpha/2}}\right] \tag{10.4.5}$$

respectively. Recall that the mean and variance of the lognormal distribution are given by

$$\text{Mean} = e^{\mu+\sigma^2/2},\ \text{Variance} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \tag{10.4.6}$$

Now, the estimates of the mean and variance of the time to failure can simply be found by replacing μ and σ^2 in equation (10.4.6) by their estimators in equation (10.4.3).

PRACTICE PROBLEMS FOR SECTIONS 10.3 AND 10.4

1. The time to failure (in months) of a component in a supercomputer is appropriately represented by an exponential distribution. Develop a sequential test for testing a hypothesis H_0: $\mu = \mu_0 = 120$ versus H_1: $\mu = \mu_1 = 72$, using (a) $\alpha = 0.05, \beta = 0.05$; (b) $\alpha = 0.05, \beta = 0.10$.
2. Referring to Problem 1, develop a sequential test for testing a hypothesis H_0: $R(100) = 0.95$ versus H_1: $R(100) = 0.80$, using (a) $\alpha = 0.05, \beta = 0.10$; (b) $\alpha = 0.10, \beta = 0.10$.
3. Suppose that a random sample of 15 hard drives is placed on life test and that the test was concluded after the seventh failure. The recorded times to failure are 3056, 3245, 3375, 3450, 3524, 3875, and 4380 hours. Estimate the mean time between failures.
4. Refer to Problem 3. (a) Estimate the hazard rate. (b) Find a 99% confidence interval for the mean time between failures. (c) Find a 99% confidence interval for the hazard rate. (d) Estimate the mean time to failure with 95% confidence.
5. In Problem 3, suppose that it was decided to complete the life test at time 4500 hours, so that the seven failures recorded in Problem 3 occurred within that time. (a) Estimate the mean time between failures. (b) Estimate the reliability at $t = 8000$ hours. (c) Estimate the hazard rate. (d) Find a 95% confidence interval for the mean time between failures. (e) Find a 95% confidence interval for the hazard rate. (f) Estimate the mean time to failure with 95% confidence.
6. Suppose that the time to failure in Problem 3 is modeled by the Weibull distribution. Using MINITAB, find the least-squares and MLE estimates of the mean and the standard deviation for both censored data and uncensored data.
7. Suppose that the time to failure in Problem 3 is modeled by the lognormal distribution. Using MINITAB, find the least-squares and MLE estimates of the mean and the standard deviation for both censored data and uncensored data.
8. Refer to Problems 3, 6, and 7, to construct probability plots (for censored data) using MINITAB for the Weibull, lognormal, and exponential distributions, and then use the Anderson–Darling criterion to decide which distribution is the best fit to these data.
9. Suppose that the time to failure of a part in an airbus gas turbine is lognormal with parameters μ and σ^2. A random sample of 10 parts is placed on life test, and the recorded times to failure are 46, 49, 54, 58, 60, 64, 66, 69, 72, and 78 months. Determine the MLE of μ and σ^2, and then find estimates of the mean and variance of time to failure of the part under investigation.

10.5 CASE STUDIES

Case Study 1 (*Failure rates in electrical cable insulation*)[1] Stone and Lawless (1979) point out that engineering experience suggests that failure voltages for two types of cable are adequately presented by Weibull models with a common shape parameter. The data in Table 10.5.1 give the failure voltage (in kilovolts per millimeter) for 20 specimens of two types of cable (data used with permission).

[1] Source: Lawless (2003) (Data used with permission).

Table 10.5.1 Failure voltage (kV/mm) for two types of cable.

Type I cable	32	35.4	36.2	39.8	41.2	43.3	45.5	46	46.2	46.4
	46.5	46.8	47.3	47.3	47.6	49.2	50.4	50.9	52.4	56.3
Type II cable	39.4	45.3	49.2	49.4	51.3	52	53.2	53.2	54.9	55.5
	57.1	57.2	57.5	59.2	61	62.4	63.8	64.3	67.3	67.7

(a) Using MINITAB, R, or JMP, fit Weibull models and examine if the least-squares and maximum likelihood estimates of the shape parameter are similar.
(b) The scale parameter α of the Weibull model has the interpretation that it is the quantile at which 63.2% of the units fail. Now determine the 63rd percentile for each data set and examine if it matches with the corresponding estimates of the scale parameter that you determined in (a).
(c) Use the estimates of the shape parameter in (a) to conclude whether the hazard rate function is increasing, decreasing, or constant.
(d) Determine the reliability of each cable at 55 kV/mm.
(e) Determine 95% confidence intervals for the scale and the shape parameters.

Case Study 2 (*Failure rate in microcircuits*)[2] Failures occur in microcircuits because of the movements of atoms in the conductors in the circuit, a phenomenon referred to as electromigration. The data in Table 10.5.2 give the failure time (in hours) from an accelerated life test of 59 conductors.

Use one of the statistical packages discussed in this book to fit the data in Table 10.5.2 to the exponential, lognormal, and Weibull models and estimate the corresponding parameters. Determine which model fits better to these data and then find the hazard rate function.

Case Study 3 (*Survival time of male mice*)[3] The data in Table 10.5.3 give the survival times (in weeks) of 208 male mice exposed to a 240-R dose of gamma radiation. Fit the data in Table 10.5.3 to the lognormal and the Weibull models and estimate the corresponding parameters. Determine which model fits these data better and then find the hazard rate function.

Table 10.5.2 Failure time (in hours) in an accelerated life test of conductors.

6.545	9.289	7.543	6.956	6.492	5.459	8.120	4.706
8.687	2.997	8.591	6.129	11.038	5.381	6.958	4.288
6.522	4.137	7.459	7.495	6.573	6.538	5.589	6.087
5.807	6.725	8.532	9.663	6.369	7.124	8.336	9.218
7.945	6.869	6.352	4.700	6.948	9.254	5.009	7.489
7.398	6.033	10.092	7.496	4.531	7.974	8.799	7.683
7.224	7.365	6.923	5.640	5.434	7.937	6.515	6.476
6.071	10.491	5.923					

[2] Source: Data Table 10.5.2 from Nelson and Doganaksoy (1995) and Lawless (2003).

[3] Source: Data Table 10.5.3 from Furth, Upton, and Kimball (1959), Kimball (1960), and Lawless (2003).

Table 10.5.3 Survival times (in weeks) of male mice.

40	48	50	54	56	59	62	63	67	67	69	70
71	73	73	76	77	80	81	81	82	83	84	86
86	87	88	88	88	88	88	89	90	90	91	93
94	95	96	97	97	98	99	99	100	100	100	100
101	101	101	102	102	103	103	103	103	103	104	104
104	105	105	106	106	106	107	108	109	109	110	110
110	111	111	111	112	113	113	114	114	115	116	116
117	118	118	118	119	119	120	120	120	121	121	123
123	124	124	124	125	125	126	126	126	126	126	127
127	127	127	128	128	128	128	129	129	129	129	129
129	130	130	130	130	131	131	132	133	133	133	134
134	134	134	135	135	135	136	136	136	136	137	137
137	138	139	139	140	140	141	141	141	141	141	142
144	144	144	144	144	145	145	146	146	146	146	147
147	147	147	148	148	148	148	149	150	151	151	151
151	152	152	153	155	156	157	158	158	160	161	162
162	163	163	164	165	165	166	168	169	171	171	172
172	174	177	177								

10.6 USING JMP

This section is not included in the book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. To meet contract specifications, the reliability of a certain type of electrical relay must be at least 0.90 at 1000 days of service. Given that the distribution of time to failure is exponential, what is the MTBF of the relays? What is the hazard rate?

2. The life distribution of certain items is $N(\mu, \sigma^2)$ with $\mu = 65$ and $\sigma^2 = 100$. Determine the force of mortality (or hazard rate) of an item at times $\mu \pm \sigma, \mu, \mu \pm 2\sigma$.

3. Referring to the situation discussed in Section 10.3, determine the maximum likelihood estimate of λ, under the assumption of exponential times, that is $f(t) = \lambda \exp(-\lambda t)$, where $\mu = 1/\lambda$.

4. Given the hazard function $h(t) = \beta t$, (a) determine the appropriate p.d.f. for $f(t)$ and its reliability function $R(t)$; (b) determine $f(t)$ and $R(t)$ when $h(t) = \beta_0 + \beta_1 t$.

5. A sample of $n = 12$ radio receivers is placed on a life test. The times to failure, reported after the receivers are on test for 3000 hours, are 1854, 2110, 2270, and 2695. Given that the hazard rate for these items is a constant, estimate their reliability at $t = 4000$.

6. In Problem 5 above, suppose that the test was stopped at the time of the fourth failure. Estimate, with 95% confidence, the reliability at time $t = 4000$ hours.

7. A sample of $n = 12$ light bulbs are left on test until the fifth bulb has failed. The recorded times to failure in hours are 1284, 1352, 1397, 1473, and 1530.
   (a) Construct the 95% confidence limits for the mean time between failure (MTBF), and then estimate the MTBF with 95% confidence.
   (b) What is the estimated reliability at $t = 2000$ hours? What is the estimated reliability at $t = 2000$ with 95% confidence?
   (c) At what time t will $R(t) = 0.5$?

8. Determine the hazard function for the gamma distribution

$$f(t|\beta, \lambda) = (\lambda^{\beta}/\Gamma(\beta))t^{\beta-1}e^{-\lambda t}$$

9. Develop a sequential sampling plan to test the hypothesis H_0: $\mu_0 = 3000$ and H_1: $\mu_1 = 2500$ with α and β risks 0.05 and 0.10, respectively, when the measured response time to failure of the test items are normally distributed with mean μ and standard deviation 225 (a) when items are replaced on failure; (b) when items are not replaced on failure.

10. Consider the following times to failure in hundreds of hours of ten systems placed in service under same conditions. The experimenter decided to terminate the experiment after 2900 hours:

15	17	20	21	23	26	28	29+	29+	29+

Fit an exponential model to these data. Prepare the probability, reliability (survival), and hazard plots (note that 29+ means that the item had life greater than 2900 hours).

11. Use the data in Problem 10 to fit Weibull and lognormal models. Construct for both models the probability, reliability (survival), and hazard plots.

12. The following data give the survival times in months of 20 patients after their brain tumors were removed:

10	17	77	27	14	17	17	17	3	2	20	9	18	7	15	7	24	56	30	16

Fit the lognormal model to this data using MINITAB or JMP, and then construct the probability, survival, and hazard plots.

13. Refer to Problem 12. Use MINITAB or JMP to find a 95% confidence interval for the mean.

14. Refer to Problem 12. Suppose that the recording was terminated after the death of the 13th patient at time $t = 17$. Use MINITAB or JMP to fit Weibull and lognormal models. Construct the probability, reliability (survival), and hazard plots for both models.

Chapter 11

ON DATA MINING

The focus of this chapter is to introduce basic concepts and tools in Data Mining.

Topics Covered

- What is data mining and big data?
- Data reduction and visualization
- Data preparation
- Classification
- Classification and regression trees

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the need for data mining.
- Learn data reduction techniques.
- Use data visualization and data preparation methods.
- Classify data using appropriate tools.
- Construct both classification and regression trees.
- Use the R software to perform classification and regression tree analyses.

11.1 INTRODUCTION

Previously, we studied a variety of probability and statistics topics starting from the very basic level suitable for those with no previous knowledge up to an advanced level of various techniques of statistics. Specifically, in Chapter 2, we learnt to divide data into different

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

classes or classify data into various classes, derive various descriptive statistics such as mean, variance, standard deviation, and others. We also learned to identify outliers that may be present in a set of data. Afterward, we focused on various distributions, in particular, the normal distribution, where we learnt how to standardize random variables. This led to a discussion on estimation and testing of hypothesis and some basics concepts in reliability theory. In Chapters 15 and 16, we will discuss regression analysis and logistic regression, where we will learn how to fit prediction models using one or more independent variables.

We dedicate this chapter to the rather popular and very useful area in statistics and computer sciences called "data mining." All the concepts or techniques mentioned previously become, one way or the other, the basis of various techniques used in the study of data mining. Therefore, throughout this chapter, we may recall some of the topics from the previous chapters. In summary, this chapter presents a general overview of data mining with the intent of motivating readers to explore the topics in data mining further, and gain expertise in this very important area.

11.2 WHAT IS DATA MINING?

Data mining is a methodology of extracting information from "big data sets or large data sets" and includes gaining knowledge using various techniques of statistics and machine learning. Machine learning is a subfield of artificial intelligence based on the concept that systems can learn from data, finding hidden patterns in big data sets and arrive at decisions with least human input. Data mining may also be described as discovering patterns and knowledge from various databases by using various techniques of data analysis and analytics, which involves database, data management, data preprocessing, or data reduction. Data reduction involves data complexity considerations, data representations of the data mining model, and inference considerations. Finally, it involves the postprocessing of discovered patterns and visualization with online updating.

As mentioned earlier, knowledge discovery in large data sets is achieved in various stages. That is data mining involves data selection, data preprocessing or data reduction, transformation, interpretation, and evaluation. The goal of data mining may be described as prediction, i.e., developing data mining prediction models, classification, clustering (see Chapter 12), and possible finding of some other functions.

We remark that the data reduction process in big data is an essential and integral part of data mining. We will discuss some of the steps that occur in the process of data reduction later in this chapter. The knowledge gained from a big data set may be used for many applications such as follows: (i) retaining customers, (ii) DNA studies, (iii) market analysis, (iv) space exploration, (v) tracking fraudulent activities, and (vi) corporate analysis and risk management.

11.2.1 Big Data

The modern advancement in computer technology, remote sensing and wireless sensors networks have made the collection of data sets very simple and cheap. Moreover, the growth and digitalization of global information storage capacity has increased and is increasing at a phenomenal rate. These two aspects have almost made the collection of very large and complex data a routine task in many areas of application.

Big data sets are very large and complex, so that traditional data-processing software cannot handle and analyze them. Some of the challenges include capturing data, data storage, and data analysis. There are also various conceptual problems associated with big data, which include volume of the generated and stored data, variety (type and nature of data), velocity at which the data is generated, how much redundancy or noise is present in the data, and how much value or knowledge is present in the data. More recently, the term "big data" is used when we use predictive analytics and other advanced analytic methods to extract value from the data. As previously mentioned, with the advancement in computer technology, remote sensing, and wireless sensors networks, data size has grown very rapidly. On the other hand, Internet of Things (IoT) has enormous amount of data. Based on an International Data Corporation (IDC) report prediction, the global data volume will grow from 33 ZB (44×2^{70}) in 2018 to 175 ZB by 2025 (Reinsel et al., 2018). Again, based on an IDC forecast, worldwide revenues for big data and business data analytics (BDA) solutions will reach $260 billion in 2022 with a compound annual growth rate (CAGR) of 11.9% over the 2017–2022 forecast period.

The desktop computers and software packages to visualize data often have difficulty handling big data. The work may require "massively parallel software" running on tens, hundreds, or even more servers. What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

When dealing with many variables, sometimes we are faced with the problem of "multicollinearity,", that is some of the independent variables are related with each other. The consequences of multicollinearity in multiple linear regression will be discussed in Chapter 16. However, Multicollinearity occurs in big data, where we have hundreds or sometimes even thousands of variable (features), and it is very common that some of the features may be related to each other. This is just one problem, and it may be associated with another broader issue known as the *curse of dimensionality*, where the data volume increases exponentially, and data become sparse as the dimensionality of data increases. However, since the big data is so complex, it may have many such problems. Therefore, before analyzing big data, it is very important to eliminate, or at least minimize the number of attributes with the data set. This process of eliminating attributes from a data set is referred to as *data reduction*. It is important to note that using data reduction techniques on the preprocessed or prepared data for data mining, is questionable, in the sense that using data reduction techniques on preprocessed data makes it more complex, and this takes more computing time, so that there is little or nothing to gain. We would recommend to an analyst that he/she maintains the right flavor between simplicity versus accuracy in the data mining process. The following section offers a brief discussion on data reduction.

11.3 DATA REDUCTION

The data mining techniques are used for extracting patterns and values from a big data set. The data reduction is important to improve the quality of the data. The good quality data obtained using data mining techniques will not only allow us to determine whether the desired problem is solvable or not but also it will lead us to obtain useful and valuable models while lessening the deleterious effects of curse of dimensionality. As such, data reduction is an integral part of data mining.

In the process of data reduction and preparing the data for data mining, it is important to know exactly what we gain and/or lose with these actions. Kantardzic (2011) notes three important points which should be kept in mind while reducing the data.

1. *Computing time*: As a result of data reduction, we expect that the data is simplified which would lead to a reduction in the time taken for data mining. In many situations, we may not be able to spend too much time on the data preprocessing phases, including a reduction of data dimensions; although, the more time we spend in preparation of the data, the better the outcome would be, i.e., we would have more valuable data mining models.
2. *Predictive/descriptive accuracy*: This is the dominant measure for data-mining models since it measures how well the data are summarized and generalized into the model. We generally would expect that by using only relevant features, a data-mining algorithm can learn not only faster but also with higher accuracy. Irrelevant data may mislead a learning process and the final model, while redundant data may complicate the task of learning and cause unexpected data-mining results.
3. *Representation of the data mining model*: The simplicity of representation of the data mining model, usually obtained with data reduction, often implies that a model can be better understood. Furthermore, the simplicity of the induced model and other results depends on its representation. Therefore, if the simplicity of representation improves, a small decrease in accuracy may be acceptable. However, the need for a balanced view between accuracy and simplicity is necessary, and dimensionality reduction is one of the mechanisms for obtaining this balance.

It would be ideal if we could achieve all three, that is reduced computer time, improved accuracy, and simplified representation at the same time, using dimensionality reduction. However, very often it happens that we gain something but lose other features. Thus, it is very important that we attain some balance in such a way that we do not lose much in what is more important to our application. Further, it is well known that no single data-reduction technique is well suited for all applications. To decide which data-reduction technique should be used depends on how much knowledge we have about the available data, such as, for example, feature correlation, noise data, outliers, relevant data, and so on. Here, we briefly discuss some of data reduction methods.

1. *Dimensionality reduction*:
   The big data are usually multidimensional and very complex. For instance, consider a regular customer database where we have a very large number of raw attributes such as in store purchases, demographic information, online purchases, social media sites visited, traveling information, nature of the use of mobile devices, nature of the responses to different surveys, number of vehicles owned, etc. This would result possibly hundreds perhaps thousands of raw attributes per customer. This would be the case where the big data set comes from chain stores with millions of customers. Working with such a huge data base, say specifically when model building is challenging due to both the computational aspect and interpretability of results, makes itself a case for reduction. Now in data reduction, one of the goals is reducing the dimensionality of the big data. One possible basic approach to reduce dimension is discarding some of the features (variables) of the big data. For instance, in multiple linear regression analysis, the dimensionality reduction is accomplished by detecting and removing irrelevant attributes that are not required for data analysis,

i.e., removing variables from the data without losing any pertinent information (see Chapter 16). A simple method is to determine pairwise correlations to detect variables or sets of variables, which may produce similar information. This way, some of the variables that provide similar information can be removed. On the other hand, if the number of features is too large, then the available data may sometimes be insufficient for data mining. In such cases, we may have to get rid of whole lot of features to obtain any valuable model.

However, dimensionality reduction can be achieved by *encoding* or *transforming* data into a more "compressed" representation while retaining almost all the information of the original data. *Principal component analysis* (PCA) and *wavelet transformation* (WT) methods are commonly used for this purpose. These methods cannot retain 100% of the original information while reducing dimension. As such, these are called as *lossy* dimensionality reduction methods. For instance, using the PCA method, one can represent ordinal data using a few mutually orthogonal (new) set of variables which are linear combinations of the original features. The weights or coefficients of the linear combinations are calculated using eigenvector analysis on the correlation matrix of the original data. On the other hand, WT transforms an original feature vector into a different vector where one can perform further data reduction on this new vector. A detailed theoretical explanation is beyond the scope of this book.

2. *Data aggregation*:
Aggregation is a commonly used statistical technique in descriptive statistics where we summarize the data so that it reduces the data volume while retaining the information needed for further analysis. For example, in the aforementioned customer data base scenario, we can aggregate monthly customer data over seasons or quarters and store in a relatively smaller volume. This can further be enhanced if we summarize data by customers' counties or geographical regions. Sometimes we refer to this method as *data cube aggregation* as we summarize multivariate (three or higher dimensional) data in cubes.
3. *Attribute selection*:
To predict a dependent variable, we may consider several independent variables, where only a few of them may be important, and the remaining independent variables may not be useful in predicting the dependent variable. Similarly, in practice, data mining problems are characterized by high-dimension data, where not all features are important, that is many of the features may be irrelevant and thus may be merely noise-producing attributes. This makes the data mining procedure less effective or may not provide very valuable and powerful models. The presence of irrelevant features in a large data set also effects adversely the important points mentioned previously, that is *Computing Time*, *Predictive/Descriptive accuracy*, and *Representation of the Data Mining Model*. Therefore, elimination of irrelevant features is an important part of the data reduction process. As we will notice in Chapter 16, quite often in regression analysis, we are seeking best sets of independent variables, using certain criteria such as *R-Sq*(*adj*), Mellows C_p, *PRESS* statistic, and *R-Sq*(*pred*). Similarly, in data-mining procedures, we like to determine important attributes by eliminating some of the irrelevant features, which results in less data, and consequently, the data-mining algorithm can learn faster, provides higher accuracy, and yields simple results which are easier to understand and apply. There are a few common methods applied for reducing the number of attributes. For instance, *Stepwise forward and/or backward selection* procedures in regression analysis help us to find

the most appropriate set of attributes from a set of attributes. The decision tree approach which we discuss in Section 11.7 is also a technique that one can use to eliminate nuisance attributes.

The other data reduction method includes *Numerosity reduction*, where we reduce the data volume by using some alternative data representations, which includes regression models (e.g., logistic regression models), some graphical methods (e.g., histograms, box-plots), and clustering methods (see Chapter 12). We can also use *Discretization* techniques where we replace continuous variables with a range of numerical values with interval-based labeled variable. For detailed discussions of these methods, readers are referred to Kantardzic (2011) and Han et al. (2011).

11.4 DATA VISUALIZATION

Another tool for preprocessing the data is what is called "data visualization." In earlier chapters, we studied several concepts of simplifying the data, such as data classification, graphical displays, numerical summaries, data tabulations, and other techniques. All these techniques form a major but not exhaustive part of the data visualization process. The data visualization techniques' such as a bubble chart, parallel coordinate plots, tree maps, GIS (Geographic Information System) charts, Data Dashboards, Key Performance Indicator, and others play a vital role in the data mining process. Some of these visualization techniques will be discussed here, but for further information, the readers may refer to Camm et al. (2014), Kantardzic (2011), and some other references given at the end of this book.

The aforementioned techniques and some others help us to visualize the data in a very simple format, so that we can easily see some of the characteristics of the data, such as range of the data, rate of frequency of different observations or any trends present in the data, and so on. This allows us to eliminate any outliers or any redundancy which may be present in the data. The visualization of the data allows the managers to see in summarized form the company's operation and communicate with the data-mining analysts about what results they are seeking. Then, the analysts may apply the appropriate data-mining techniques and get the results in simple form so that they can be interpreted easily and that the valuable information can be extracted from them.

In data visualization, the primary goal of data presentation, whether it is in a tabular form or in graphical form, should be such that the tables and charts should be free of unnecessary labels or grid lines etc. In the following example, we are given the number of laptop computers manufactured by a high-tech company per day during the month of June of a given year.

Example 11.4.1 (Data visualization) *The following data give the number of laptop computers manufactures by a high-tech company during the month of June (30-day period) in a given year*

445, 460, 428, 464, 479, 438, 428, 441, 448, 420, 434, 453, 419, 430, 456
468, 476, 449, 426, 478, 410, 456, 440, 428, 455, 463, 437, 446, 421, 483

Exhibit these data in various formats and explain which format might be more informative through data visualization.

Solution: We write the above data in a couple of different formats, say, tabular and graphical formats, and note which format would be more informative with minimum efforts when we use the data visualization technique. Tufte (2001) in his book "The Visual Display of Quantitative Information" describes that one of the novel ideas to evaluate the data visualization technique is data-ink ratio. The data-ink ratio measures the proportion of how Tufte describes "data-ink" namely, as the total amount of ink used to present a set of data in a tabular or graphical format.

Compare to Table 11.4.1 and Figure 11.4.1a, clearly Table 11.4.2 and Figure 11.4.1b improves the visibility of important information in the data due to the visualization technique used. In general, as we can see in the above examples, graphs provide information faster, clearer, and easier for the readers to grasp the information contained in the given

Table 11.4.1 Table format of the data with grid-lines (low data-ink ratio).

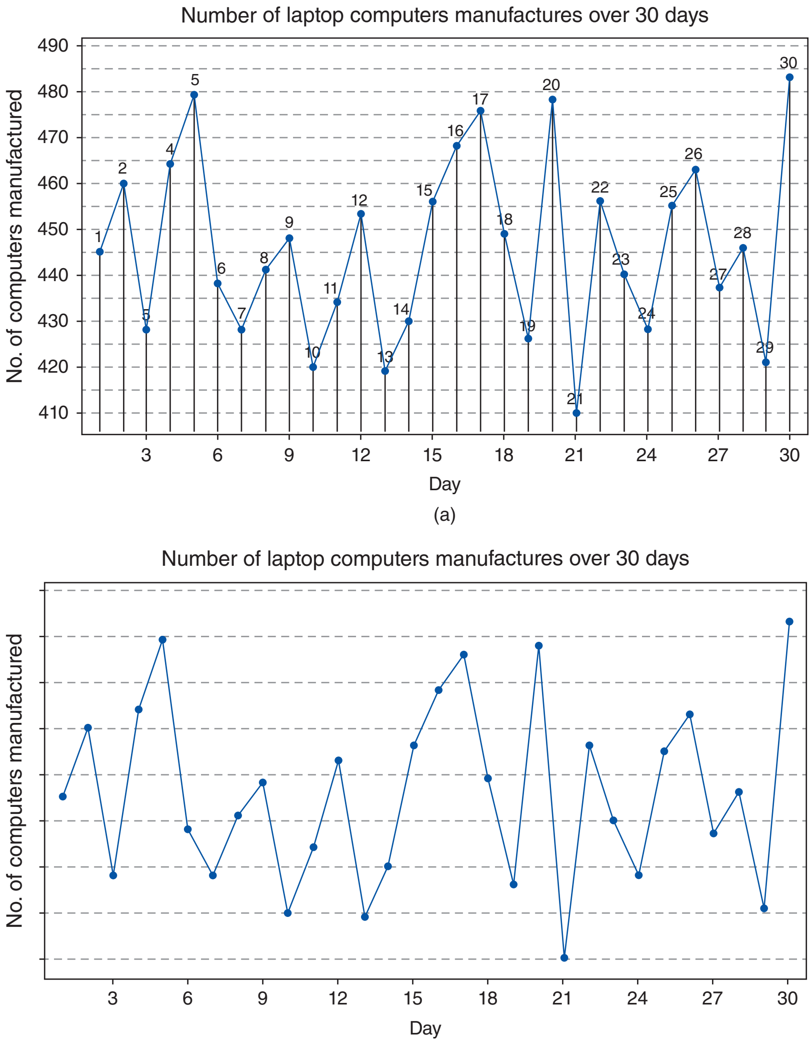
Day	No. of computers manufactured	Day	No. of computers manufactured	Day	No. of computers manufactured
1	445	11	434	21	410
2	460	12	453	22	456
3	428	13	419	23	440
4	464	14	430	24	428
5	479	15	456	25	455
6	438	16	468	26	463
7	428	17	476	27	437
8	441	18	449	28	446
9	448	19	426	29	421
10	420	20	478	30	483

Data represent the number of laptop computers manufactured by a high-tech company over 30-day period.

Table 11.4.2 Table format of the data without grid-lines (high data-ink ratio).

Day	No. of computers manufactured	Day	No. of computers manufactured	Day	No. of computers manufactured
1	445	11	434	21	410
2	460	12	453	22	456
3	428	13	419	23	440
4	464	14	430	24	428
5	479	15	456	25	455
6	438	16	468	26	463
7	428	17	476	27	437
8	441	18	449	28	446
9	448	19	426	29	421
10	420	20	478	30	483

Data represent the number of laptop computers manufactured by a high-tech company over 30-day period.



(a)

(b)

Figure 11.4.1 (a) Graphical format of the above data with projected lines (low data-ink ratio). (b) Graphical format of the above data without projected lines (high data-ink ratio).

data set. For lack of space, we will not discuss the data visualization technique any further. However, the reader may find many various kinds of tables and graphs discussed in this book are very helpful in data visualization.

Example 11.4.2 (Iris Flower Data by Fisher (1936) and Anderson (1935)) *Use the popular Iris flower data set to exhibit various visualisation techniques using R software. Table 11.4.3 shows the first few observations of the data set, but the complete data set is available on the website: www.wiley.com/college/gupta/statistics2e. The "Species" is a categorical variable with three classes (Setosa, Versicolor, and Virginica) and "Sepal Length," "Sepal Width," "Petal.Length," and "Petal Width" describe numerical measurements of the flowers as shown in Figure 11.4.2.*

Table 11.4.3 Iris flower data (only the first few observations are shown here).

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.3	3.0	1.1	0.1	Setosa
4.4	2.9	1.4	0.2	Setosa
4.4	3.0	1.3	0.2	Setosa
4.5	2.3	1.3	0.3	Setosa
4.6	3.4	1.4	0.3	Setosa
4.6	3.2	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.7	3.2	1.6	0.2	Setosa
4.8	3.4	1.6	0.2	Setosa
4.8	3.0	1.4	0.1	Setosa
4.8	3.4	1.9	0.2	Setosa
4.8	3.1	1.6	0.2	Setosa
4.8	3.0	1.4	0.3	Setosa
4.9	3.1	1.5	0.2	Setosa
4.9	2.4	3.3	1.0	Versicolor
4.9	2.5	4.5	1.7	Virginica

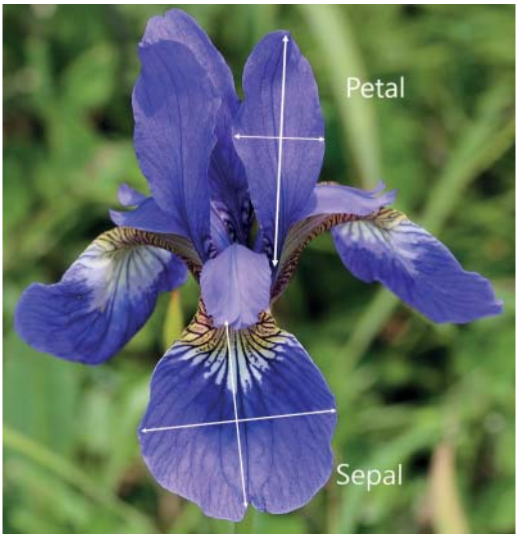


Figure 11.4.2 Iris versicolor.

USING R

First, we use some built in R functions to visualize numerical summaries of these Iris data as follows. The syntax “iris[7:10,]” is used to visualize Observations 7–10 in the Iris data frame. The “summary()” and “quantile()” functions provide the summary statistics and five number summary, respectively. The corresponding outputs are shown just after the following R code.

```
#Load iris data from R 'datasets' library
attach(iris)

#Visualise Observations 7-10
iris[7:10,]

#R output
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
7	4.60	3.40	1.40	0.30	setosa
8	5.00	3.40	1.50	0.20	setosa
9	4.40	2.90	1.40	0.20	setosa
10	4.90	3.10	1.50	0.10	setosa

```
#Summary of attributes
summary(iris)

#R summary output
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	Min.: 4.300	Min.: 2.000	Min.: 1.000	Min.: 0.100	setosa: 50
2	1st Qu.: 5.100	1st Qu.: 2.800	1st Qu.: 1.600	1st Qu.: 0.300	versicolor: 50
3	Median: 5.800	Median: 3.000	Median: 4.350	Median: 1.300	virginica: 50
4	Mean: 5.843	Mean: 3.057	Mean: 3.758	Mean: 1.199	
5	3rd Qu.: 6.400	3rd Qu.: 3.300	3rd Qu.: 5.100	3rd Qu.: 1.800	
6	Max.: 7.900	Max.: 4.400	Max.: 6.900	Max.: 2.500	

```
#Quantiles of the first variable in the data frame
quantile(iris[,1])

#R summary output
```

0%	25%	50%	75%	100%
4.3	5.1	5.8	6.4	7.9

The following R code can be used to obtain basic graphical displays such as histogram, boxplot, and scatter plots. As shown in Figure 11.4.3d, the scatter plot with groups helps us not only to understand the trend between sepal width versus sepal length but also allows to see different trends for each flower type with some clustering structures.

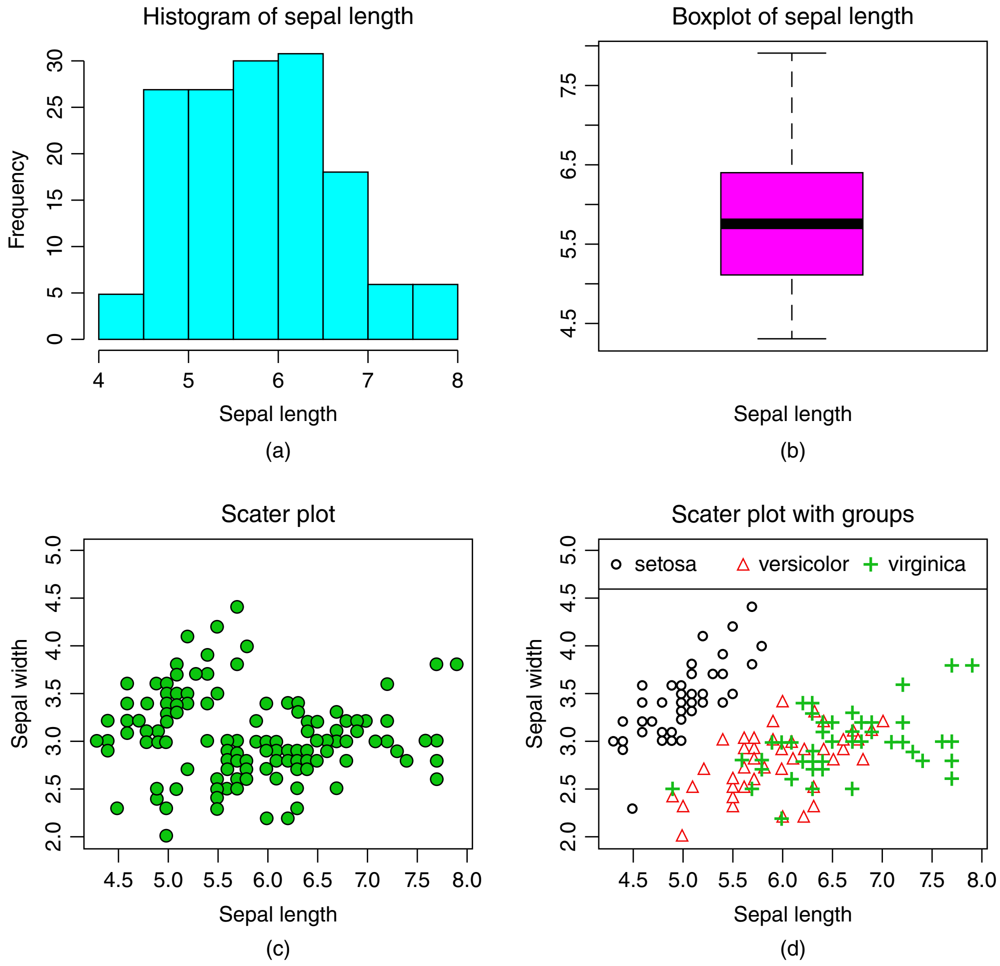


Figure 11.4.3 Various graphical visualizations for Iris data.

```
par(mfrow=c(2,2))
#Histogram of Sepal Length
hist(iris[,1], main=“Histogram of Sepal Length”, xlab =“Sepal Length”, lwd=2, col=5)

#Boxplot of Sepal Length
boxplot(iris[,1], main=“Boxplot of Sepal Length”, xlab =“Sepal Length”, lwd=2, col = 6)

#Bivariate scatter plot of Sepal Length vs Sepal Width
plot(iris[,1], iris[,2], main=“Scater plot”, xlab =“Sepal Length”,
ylab =“Sepal Width”, cex=2, lwd=2, col=3, pch=20, ylim=c(2,5))
```

```
#Bivariate scatter plot of Sepal Length vs Sepal Width with flower types
plot(iris[,1], iris[,2], main=“Scater plot with groups”, xlab =“Sepal Length”,
ylab =“Sepal Width”, lwd=2, col=as.numeric(iris[,5]), pch=as.numeric(iris[,5]),
ylim=c(2,5))
legend(“top”, c(“setosa”, “versicolor”, “virginica”), col=c(1,2,3),
pch=c(1,2,3),horiz = TRUE)
```

Finally, we exhibit two advanced data visualization tools in R. The R function “scatterplot3d()” yields a very useful graphical display that permits visualization of three variables at a time. As shown in Figure 11.4.4, this 3D graph helps us to visualize the nature of the association of all three variables side by side. Further, one can add a regression or a response surface to these data to visualize the 3D trend.

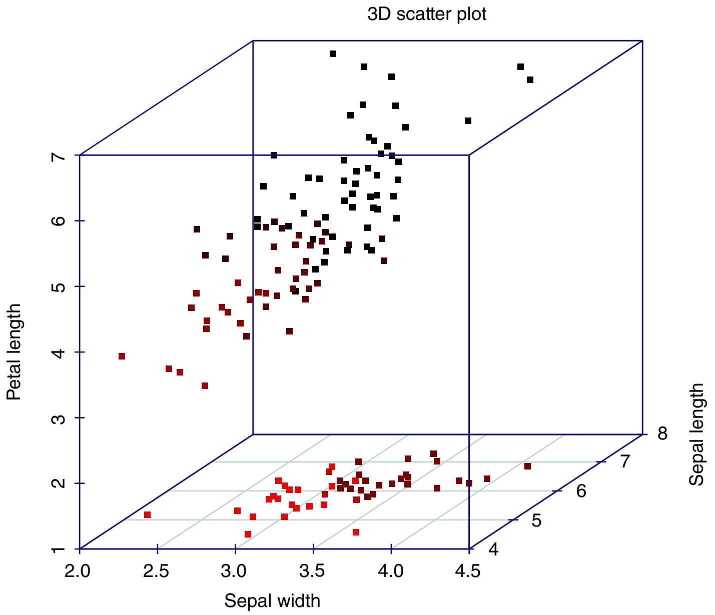


Figure 11.4.4 3D scatter plot for Iris variables “Sepal width”, “Sepal length”, and “Petal length.”

The function “ggplot()” is another highly useful R tool for data visualization. There are lot of variations and additional graphing available for this particular function. Here, we only exhibit one of the primary uses of this tool. As shown in Figure 11.4.5, we plot side-by-side boxplots that convey information about the distributions of “Petal Length” of all three flower types in our example. Figure 11.4.6 shows an *aesthetic* scatter plot where we can visualize more than two variables in a 2D plane. Aesthetic is a visual property of the objects which characterizes a property of a variable by using the size, the shape, or the color of the points on the graph. In Figure 11.4.6, we visualize attributes “Sepal Width,” “Sepal Length,” “Petal Length,” and “Species Type.”

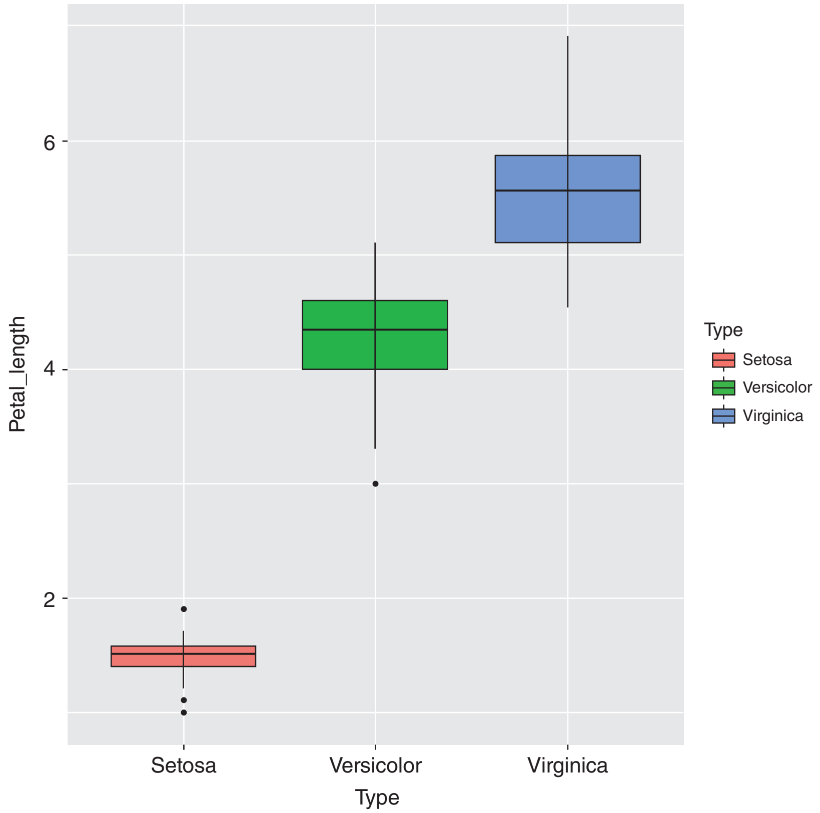


Figure 11.4.5 Side-by-side boxplots for attribute petal length for three flower types in Iris data.

USING R

The following R code can be used to obtain 3D scatter plot, side-by-side boxplots, and aesthetic scatter plot. It is required to install both “scatterplot3d” and “ggplot2” libraries to run the R code.

```
#3D scatter plot
library(scatterplot3d)
scatterplot3d(iris[,2], iris[,1], iris[,3], main =“3D scatter plot”,
xlab =“Sepal Width”, ylab =“Sepal Length”, zlab = “Petal Length”, highlight.3d=TRUE,
lwd=2, col.axis=“blue”, col.grid=“lightblue”, pch=20)


#Rename variables for plotting purpose
Type = iris[,5]
Petal_Length = iris[,3]
library(ggplot2)
```

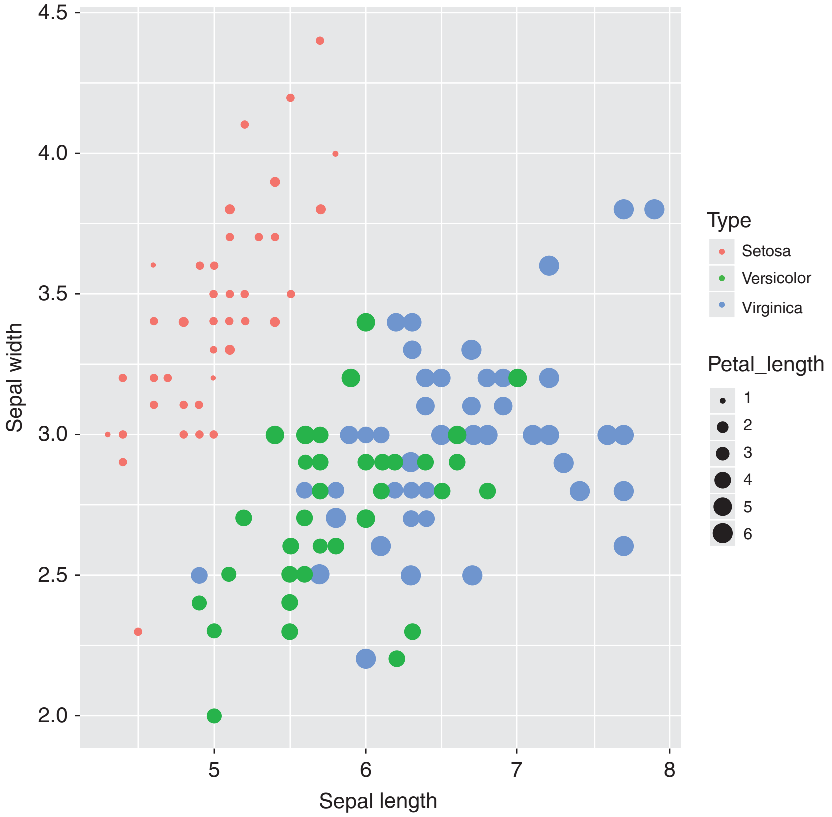


Figure 11.4.6 Aesthetic scatter plot for attributes “Sepal Width,” “Sepal Length,” “Petal Length,” and “Species Type” in Iris data. The numerical values (1–6) shown in the legend indicate that the “Petal Lengths” are proportional to the corresponding bubble sizes.

```
#Side-by-side boxplots
ggplot(iris, aes(x=Type, y=Petal_Length, fill=Type))+ geom_boxplot()


#Aesthetic scatter plot
p = ggplot(data=iris) + geom_point(mapping = aes(x = iris[,1],
y = iris[,2], color = Type, size = Petal_Length))
p + xlab(“Sepal Length”) + ylab(“Sepal Width”)
```

We also recommend readers to use another popular data visualization R package “rggobi.” The “rggobi” package provides a command-line interface to “GGobi,” an interactive and dynamic graphics package for data analysis, which provides a more robust and user friendly interface (see Wickham et al., 2008).

11.5 DATA PREPARATION

In the above sections, we gave some discussion about certain aspects of big data, including data reduction and data visualization, which constitute an important part of data preparation for data mining. In this section, we discuss some other aspects of data preparation for data mining.

11.5.1 Missing Data

In practice, even though in data mining applications we have very large data at our disposal, yet it is very rare that we have complete data sets. For example, in the study of DNA, some of the observations may be missing because during the study period, some patients may have died, some may have moved out of the area, some may have voluntarily withdrawn from the study, or some may have been diagnosed with some disease, such as cancer, HIV, or due to some other reasons. In data mining techniques, some of the techniques may provide satisfactory results by discarding observations with missing values, while other techniques may need to have complete data sets. Furthermore, as we have a very large data set, discarding some observations with missing values may still leave a significantly large data set for analysis. Thus, the data with discarded observations may still give valuable data mining models. Now, the second option is to replace the missing values with their estimated values (see Chapter 17). A third option is to use some data-mining algorithms, such as, for example, classification decision trees and continuous variable decision trees. It is generally assumed that decision trees are not seriously affected by missing data as they have their own way to handle the missing values. For example, if the predictor is categorical and has some missing values, then the tree algorithm will just consider missing values as another class. For numerical predictor with missing values, for instance, R "rpart" package uses a "surrogate" approach to estimate the missing datum using the other independent variables (see Section 11.7.1.2 for more details about "rpart" package). Data-mining analysts also use some of the following options to replace the missing observation values.

1. Replace all missing observation values with a single global constant, which is usually application dependent.
2. Replace a missing observation value with its feature mean.
3. Replace a missing observation value with its feature mean for the given class.

Note that if the missing values are replaced with a constant value or a feature mean without much justification, then they may introduce some noise in the data, which makes the data biased and thus gives us bad results.

Further note that sometimes discarding large number of observations with missing values may workout well, especially if the features or variables to be discarded are highly correlated with other variables, which have most of their observations with known values. This step would not only minimize the loss of any information but also sometimes give better information. For instance, we will notice in model building in Chapter 16 that if we discard some of the highly correlated variables, better models could be obtained.

11.5.2 Outlier Detection and Remedial Measures

In Chapter 2, we studied various plots and tables to summarize the data set. In particular, we studied boxplot (see Section 2.2.8), which helps us detect the mild outliers, and the extreme outliers via interquartile range (IQR) based (1.5 IQR and 3 IQR) rules. In general, we tend to discard all the extreme outliers unless we determine that some outlier(s) has appeared erroneously. The presence of genuine outliers in the data set can produce very lopsided results. For instance, consider the following example.

Example 11.5.1 (Outlier detection) *The following data give the salaries (in thousands of dollars) of 10 randomly selected employees from a large company.*

18, 12, 11, 15, 19, 16, 29, 23, 17, 100

The company has several hundred thousand employees. We are interested in finding the average salary of an employee in that company.

Solution: Suppose we find the sample mean to determine the average salary. So that, we have

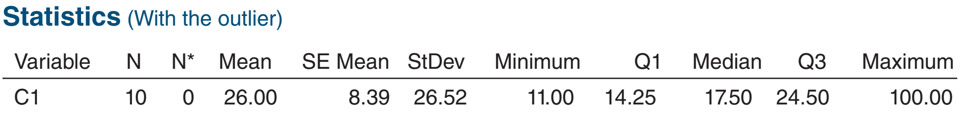
$$\bar{X} = \frac{1}{10}(18 + 12 + 11 + 15 + 19 + 16 + 29 + 23 + 17 + 100) = 26.0$$

This shows that the average salary of an employee in that company $260,000. This clearly is not a valid result and this is because of the observation 100, which may be the salary of the CEO of the company. The observation 100 is not there erroneously, but as it shows in the boxplot in Figure 11.5.1, it is a genuine extreme outlier. Now, we find the sample mean without this outlier, which is given by

$$\bar{X} = \frac{1}{9}(18 + 12 + 11 + 15 + 19 + 16 + 29 + 23 + 17) = 17.78$$

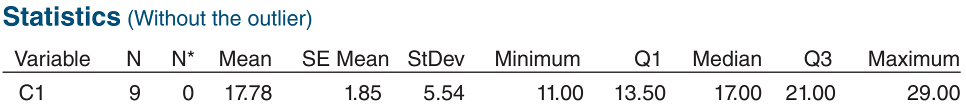
Thus, the average salary without the outlier turns out to be 17.78, that is the average salary of an employee in the company is $177,800. This clearly seems to be a valid result. We see here that how the extreme outlier(s) in the data can adversely affect the final results. Therefore, an appropriate solution for extreme outliers is that they should always be discarded.

However, as it shows in the following calculations, the sample quartiles (Q1, median, and Q3) seem to be very resistant to the extreme outlier. Therefore, we can report the median salaries 17.50 (with the outlier) or 17.00 (without the outlier) as the average salary.



Statistics (With the outlier)

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C1	10	0	26.00	8.39	26.52	11.00	14.25	17.50	24.50	100.00



Statistics (Without the outlier)

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C1	9	0	17.78	1.85	5.54	11.00	13.50	17.00	21.00	29.00

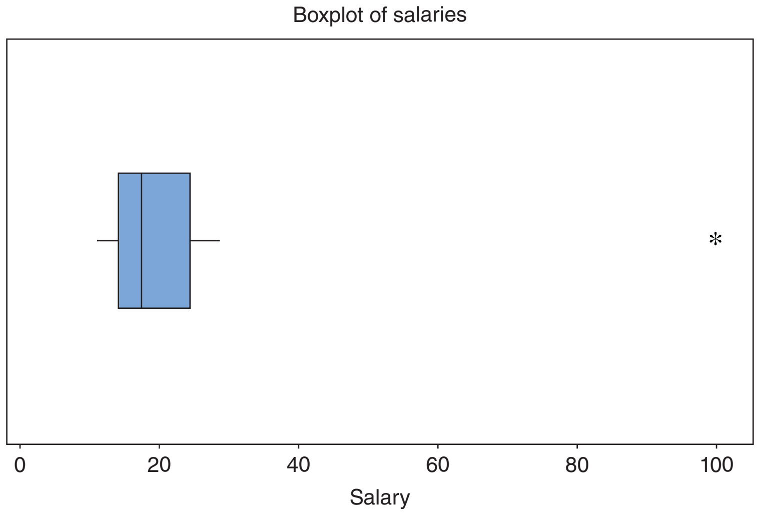


Figure 11.5.1 Boxplot of salary data. Note that the "*" represents an outlier observation.

Next, we consider the problem of how to deal with the mild outliers. The mild outliers are generally not discarded without further investigation. That is, a full investigation should be launched to determine whether these observations are due to some human errors incurred perhaps when transferring the data from one source to another source, or due to some malfunctioning of a machine, or due to any other such reason. If we detect some errors, then they either should be fixed, and replaced with correct observations, or mild outliers should be discarded.

There are some remedies existing for the cases where one simply cannot discard outliers. For instance, we can lessen the effect of the outliers by putting lower weights on the outliers. In estimation problems, one can use more robust statistics such as median absolute deviation (MAD), Huber, and Bisquare loss functions instead of the squared error loss function (see Huber, 2011; Jayalath and Gunst, 2017).

11.6 CLASSIFICATION

Data classification plays an important role in data mining process when building models. For example, a marketing management team may be interested in analyzing customers with big profile, median or low profile. For example, customers with a large profile may spend \$300.00 or more per trip. Customers with a median profile may spend \$100.00–300.00 per trip. Finally, customers with a low profile may spend less than \$100.00 per trip. Another example of classification is where a bank loan officer would like to identify applicants as no, low, medium, or high credit risk. In the first example, we are classifying the data into three groups while in the second example into four groups. A classification approach starts with a data set in which class labels or class assignments are known. That is, the classification methods can be used to classify any future data into groups/categories and help build valuable models. For example, in the aforementioned example of loan applicants, a classification model that predicts the credit risk can be built based on the data which have been collected on past applicants over a significant period of time. In addition to the usual credit rating, the loan officer may like to examine other attributes (or predictor variables, see Chapter 16) such as education, employment history, home ownership, savings, or investments. Note that in this example, attributes are the predictor variables, credit rating is the target or outcome, and the data on each customer constitutes an observation vector. If the target is categorical, then we use classification algorithms to analyze the data, but if the target is a continuous variable, then we use regression algorithms or regression models. Next, we briefly discuss, with the help of an example, performance measures of classification algorithms or classification models. For regression models, the readers are referred to the discussions of Chapters 15 and 16.

11.6.1 Evaluating a Classification Model

In any model building process, we expect some errors will occur. In other words, we cannot expect that prediction will give us a 100% correct value of any outcome. Similarly, a classification model is also bound to give some erroneous results. Classification errors are usually displayed in a matrix called the *confusion matrix.* A confusion matrix illustrates clearly which classifications are correct and which are incorrect. Tables 11.6.1 and 11.6.2 show the confusion matrices for two-class and three-class classification models.

Suppose we are dealing with a two-class classification (true and false, symbolically represented by 1 and 0, respectively)

Table 11.6.1 Confusion matrix for two-class classification model.

	Predicted class	
True class	0	1
0	(0,0)	(0,1)
1	(1,0)	(1,1)

where the entries (0, 0), (0, 1), (1, 0), and (1, 1) are interpreted as follows:

(0, 0) means a False-Class is predicted as False-Class (F+)
(0, 1) means a False-Class is predicted as True-Class (F−)
(1, 0) means a True Class is predicted as False-Class (T−)
(1, 1) means a True-Class is predicted as True-Class (T+)

Now, we consider a confusion matrix when we are classifying a data set into three classes (below-average, average, and above-average symbolically represented by 0, 1, and 2, respectively).

Table 11.6.2 Confusion matrix for three-class classification model.

	Predicted class		
True class	0	1	2
0	(0,0)	(0,1)	(0,2)
1	(1,0)	(1,1)	(1,2)
2	(2,0)	(2,1)	(2,2)

Where the entries (0, 0), (0, 1), (0, 2), ..., (2, 2) are interpreted as follows:

(0, 0) means an below-average class is predicted as below-average class
(0, 1) means an below-average class is predicted as average class
(0, 2) means an below-average class is predicted as above-average class
(1, 0) means an average class is predicted as below-average class
⋮
(2, 2) means an above-average class is predicted as above-average

A confusion matrix can be extended for any number of classes. A general rule is that all the diagonal entries in a confusion matrix represent the *true-classification*, and all *off diagonal* entries in a confusion matrix represent the *false-classification.* Below we consider a two-class classification study.

Pancreas is one of the major glands in the human body. It secretes many digestive juices, including insulin, which help the cells in the body use glucose, which is their main fuel. Every year more than fifty thousand Americans are diagnosed with pancreatic cancer. In the following example, we consider a study on pancreatic cancer.

Example 11.6.1 (Pancreatic Cancer Data) *In a major cancer center, 850 suspicious pancreatic cancer patients are tested using radiology diagnostic tests. The data in table given below provides the test results as whether patients have pancreatic cancer or patients do not have pancreatic cancer and symbolically they are represented by 1 and 0, respectively.*

Table 11.6.3 without row totals and column totals represents the confusion matrix of a two-class classification model. As mentioned previously, the *diagonal* entries in the confusion matrix represent the *true-classification*, and all *off diagonal* entries in the confusion matrix represent the *false-classification.* In a m-class classification model, the number of false-classifications is equal to number of entries in the confusion matrix minus the number of entries on its diagonal, $m^2 - m$. The evaluation measures of the classification model are determined using the confusion matrix.

Table 11.6.3 Confusion matrix for two-class classification model.

	Predicted class		
True class	0	1	Total
0	194	32	226
1	46	578	624
Total	240	610	850

The *overall error rate* in the classification model are defined as the percentage of misclassified observations (sum of all off diagonal entries in the confusion matrix). So that in the present example, the overall error rate in the classification model is

$$\text{Overall error rate} = \frac{32 + 46}{578 + 32 + 46 + 194} = \frac{78}{850} = 9.176\%$$

Therefore, the overall error rate of the classification model in this example is 9.176 %. The *accuracy rate* for the classification model is defined as (100 − overall error rate)%, so that in this example, the accuracy rate for the classification model is 90.824%. In addition to the overall error rate, we would like to see the error rate for each class as well. For instance, in the above example, misclassification rate or error rate in class 1 and in class 0, also called as *False Negative* and *False Positive*, respectively are given by

$$\text{Class 0 error rate} = \frac{32}{32+194} = \frac{32}{226} = 14.159\%$$

$$\text{Class 1 error rate} = \frac{46}{578+46} = \frac{46}{624} = 7.372\%$$

Other measures that are used to determine the performance of the classification model are called *specificity* and *sensitivity*, and they are defined as follows.

The *specificity* of a classification model is its ability to correctly predict Class 0 (negative) observations. Thus, the specificity in the present example is

$$\text{Specificity} = \frac{194}{226} = 1 - \text{class 0 error rate} = 1 - 0.14159 = 85.841\%$$

The *sensitivity* of a classification model is its ability to correctly predict class 1 (positive) observations. Thus, the sensitivity in the present example is

$$\text{Sensitivity} = \frac{578}{624} = 1 - \text{class 1 error rate} = 1 - 0.07372 = 92.628\%$$

We note that a confusion matrix shows the final classification of all the observations based on a classification algorithm. However, the classification algorithms usually predict the likelihood of the observations belonging to the existing classes. Then, given a certain *cutoff* value, it splits observations into its classes. As such, the cutoff value plays a vital role in classifying observations into classes. The default cutoff value in many algorithms is 0.5. The following example illustrates the role of cutoff values.

Example 11.6.2 (Challenger O-ring Data) *We will illustrate the use of cutoff values in a simple classification algorithm, logistic regression, which we will discuss in detail in Chapter 16. We will use the data on launch temperature and O-ring failure for the 24 space shuttle launches prior to the Challenger disaster of January 1986 (see Dalal et al., 1989; Draper, 1995). There are six O-rings used on the rocket motor assembly to seal field joints. Table 11.6.4 presents the launch temperatures and whether at least one O-ring failure (=1) occurred or not (=0) in that launch and the estimated probability of failure in each trial.*

Using the methods explained in Chapter 16, we estimate the fitted logistic regression model

$$\hat{\theta} = \frac{1}{1 + \exp(-(10.8753 - 0.1713 * \mathit{Temperature}))}$$

that provides the predictive probability ($\hat{\theta}$) of failure in each trial for a given temperature. Then, we use this predictive model to estimate the probability of failing at least one O-ring for given temperatures, and the results are summarized in Table 11.6.4. The predictive probability curve along with the O-ring failure data are shown in Figure 11.6.1.

To demonstrate the role of cutoff value in classification, we will pick three different cutoff values 0.70, 0.50, and 0.20 to classify O-ring failure data. For the first case, the observations classify as class 1 if the predictive probability $\hat{\theta} > 0.70$ and Class 0 otherwise.

Table 11.6.4 Challenger O-ring failure data.

Temperature		Predicted	Temperature		Predicted
($^\circ F$)	Status	probability $\hat{\theta}$	($^\circ F$)	Status	probability $\hat{\theta}$
53	1	0.86	70	1	0.25
56	1	0.78	70	1	0.25
57	1	0.75	72	0	0.19
63	0	0.52	73	0	0.16
66	0	0.39	75	0	0.12
67	0	0.35	75	1	0.12
67	0	0.35	76	0	0.10
67	0	0.35	76	0	0.10
68	0	0.32	78	0	0.08
69	0	0.28	79	0	0.07
70	0	0.25	80	0	0.06
70	1	0.25	81	0	0.05

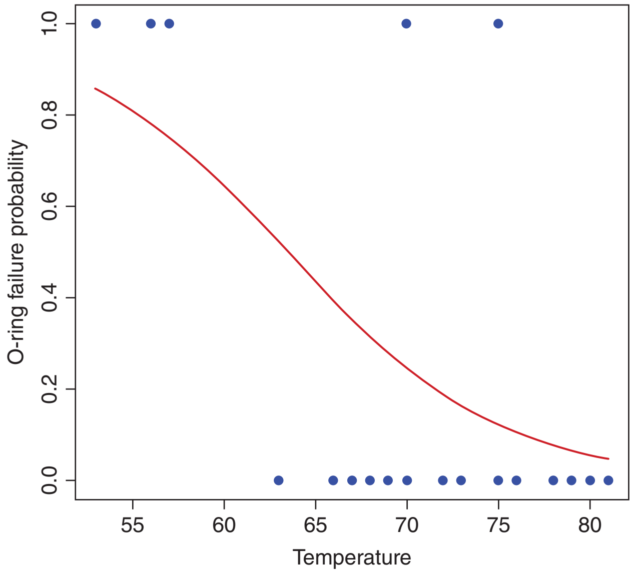


Figure 11.6.1 O-ring failure challenger data with its logistic regression based predictive probability curve.

The top panel of Table 11.6.5 shows the complete classification for this case, and it provides a higher class 1 classification error (0.57) and zero class 0 classification error. However, when we decrease the cutoff value, there are less and less class 1 misclassifications, but eventually, class 0 misclassification rate increases. This behaviour can clearly be seen in Figure 11.6.2 for a range of cutoff values we selected from 0 through 1.

Table 11.6.5 Confusion matrix for challenger data for different cutoff values.

Cutoff value = 0.70			
	Predicted class		
True class	0	1	Error proportion
0	17	0	0/17 = 0.00
1	4	3	4/7 = 0.57
Overall	21	3	4/24= 0.17
Cutoff value = 0.50			
	Predicted class		
True class	0	1	Error proportion
0	16	1	1/17 = 0.06
1	4	3	4/7 = 0.57
Overall	20	4	5/24= 0.21
Cutoff value = 0.20			
	Predicted class		
True class	0	1	Error proportion
0	9	8	8/17 = 0.47
1	1	6	1/7 = 0.14
Overall	10	14	9/24= 0.38

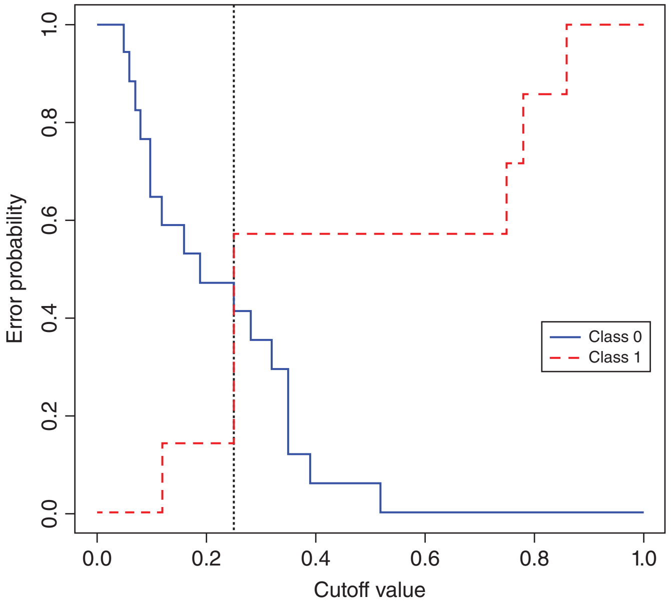


Figure 11.6.2 Classification error probability for various cutoff values.

In practice, it is important to evaluate how critical both class 1 and class 0 errors are for a given application, and one should evaluate the trade-off between errors. In this application, it is very critical to classify 1's as 0's than 0's as 1's. It would be wise in this application to allow the minimizing of a class 1 error, while allowing a manageable maximum class 0 error. For instance, the selection of the cutoff value of 0.25 provides 0.14 and 0.47 for class 1 and class 0 error probabilities, respectively, and this particular selection is indicated by a vertical line in Figure 11.6.2.

The receiver operating characteristic (ROC) curve is an alternative graphical display that can be used to investigate the quality of the classifier in a classification problem. In a ROC curve, we plot the "Sensitivity" against the "1-Specificity" for each cutoff value. For instance, when the cutoff value is zero, the class 0 and class 1 error probabilities are 1 and 0, respectively, and therefore, both the values of 1-specificity and sensitivity become 1. On the other hand, when the cutoff value is 1, both the values of 1-specificity and sensitivity become zero. A straight line simply passes through these two extreme ends may represent a random classifier (see the diagonal dashed line in Figure 11.6.3) where most classifiers fall above this line within the upper triangular area of the graph. The ROC curve of a perfect classifier should go straight up the Y-axis and then move to the right parallel to the X-axis mimicking a right triangular shape. Area under the ROC curve (AUC) score is usually considered a better measure of a classification model. In fact, AUC can be used to compare classification models in the cases where we have few models to compare. Note that the AUC score ranges from 0.5 (random classifiers) through 1 (perfect classifier). Figure 11.6.3 shows the ROC curve for the data and the model discussed in Example 11.6.2. We note that the logistic regression classifier we adapted in Example 11.6.2 performs better than a random classifier as its AUC is somewhat higher than 0.5.

In the following section, we exhibit classification data into different exhaustive regions in the presence of many predictors to classify or predict outcomes. Of course, in these

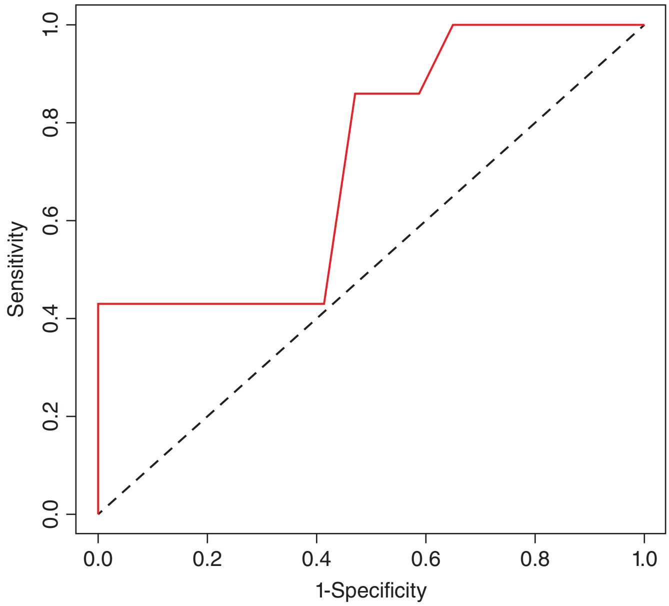


Figure 11.6.3 Receiver operating characteristic (ROC) curve for Example 11.6.2.

methods, we should repeatedly use the cutoff values to partition variables to build what we call the *Tree* structures.

11.7 DECISION TREES

A decision tree is a flowchart-like tree structure, where at each *internal node* we perform a test on an attribute using the aforementioned criteria, such as cutoff values. As a result, each branch of the tree structure represents an outcome of the test, and each branch continues to split until it terminates producing a class labeled outcome. In other words, a decision tree is a machine learning technique that recursively partitions data into smaller subsets such that data within each subset are more homogeneous or less *impure*, and data between subsets are heterogeneous. *Impurity* is a measure of heterogeneity of the outcome classes of the response variable, and as such, a good decision tree would produce information on less impurity. Figure 11.7.1 shows a decision tree structure which helps us define the basic terminologies in decision trees.

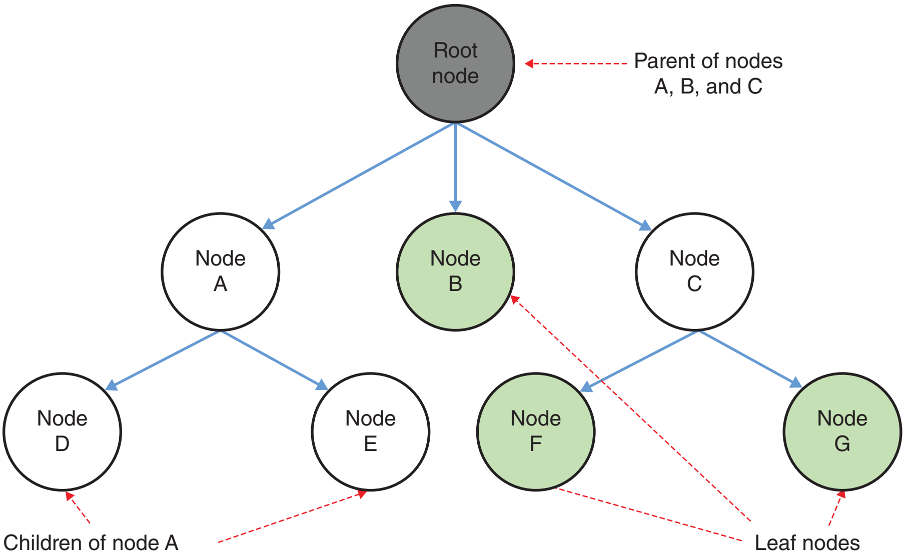


Figure 11.7.1 A decision tree structure.

The basic terminologies in decision trees.

- Parent of a node Q is the immediate predecessor node.
- Children of a node Q are the immediate successors of Q, i.e., Q is the parent node.
- Root node is the top node of the tree; the only node without parents.
- Leaf nodes are nodes which do not have children.
- A K-ary tree is a tree where each node (except for leaf nodes) has K children. When $K = 2$, it is a binary tree.
- Depth of a tree is the maximal length of a path from the root node to a leaf node.

The decision trees can be used for classification. For instance, assume that we train a decision tree using a given set of data, and now, all the decision rules (or tests) in each node are readily available. If we have a new item to classify (say $\mathbf{X}$), it is now a matter of fact how we follow all the decision rules to reach the terminal node, where we have a class label for which a new observation should belong. The decision tree technique can be applied to predict outcomes, and to identify the important predictors, such as in regression methods discussed in Chapters 15 and 16.

Unlike many of the statistical techniques which are discussed throughout this book, the decision trees require no distributional assumptions. It can also efficiently handle big data and produce meaningful results. The classification accuracy is generally higher and comparable with other statistical methods, such as logistic regression in the case where the output variable is binary. Applications of decision trees are quite common in engineering, medicine, agriculture, and finance. Iterative Dichotomiser (ID3) is one of the first decision tree algorithms developed by Quinlan (1986) and later he expanded his ID3 work to develop the C4.5 algorithm. In Breiman et al. (1984), a set of statisticians, published a monograph named "CART: Classification and Regression Trees" that provides comprehensive details of binary tree structures, algorithms, and its theories. Therefore, both ID3 and CART algorithms consider as the building blocks of its field.

11.7.1 Classification and Regression Trees (CART)

CART is primarily a binary splitting tree which provides an appealing tree-like graphical display that enables a straightforward interpretation of data. Importantly, CART can handle both quantitative and qualitative responses and predictors. In the construction of trees, the trees are grown to the maximum possible size and is then got rid of unnecessary splits by *pruning* sequentially. The pruning depends on a *cost-complexity* measure. Usually, the algorithm produces a set of pruned trees and finally selects the final tree based on its predictive ability.

Overly large trees would cause over fitting and producing unreliable results. There are two possible methods that can be used to avoid too large trees. The first, called an *Early stopping* technique, is one method that can avoid growing large trees, and results in smaller trees which saves computational time. In this method, a stopping criteria may be the number of observations in a node that undercuts the minimum number of observations. If the criterion is fulfilled, the current node will not be split any further. The second method is the aforementioned *pruning* technique. In pruning, we grow a large tree and cut afterward. The full tree is grown, and each split is examined to see if it brings a reliable improvement. There are two different ways we can examine splits. One method is called the *top-down* approach, that is starting from the first split made, and proceed to bottom layers, or the second method, called the *bottom-up* approach, that is starting at the splits above the leaf nodes and moving to the top layers. The bottom-up method is somewhat more common, as the top-down method could lead to discard a whole sub-tree due to a bad split in a top layer, though it may be that there is a lot of good splits in the subsequent layers.

Classification Trees

Let us consider a classification problem where the response variable Y is categorical with J categories or classes, and a set of predictor variables $X_1, X_2, \ldots, X_p$. We would like to predict the outcome of Y for a new observation $X^h = x_1^h, x_2^h, \ldots, x_p^h$. Now, the objective is to split the predictor space of X into J disjoint sets, $S_1, S_2, \ldots, S_J$, such that we can predict the value of Y to be i if $X^h \in S_i$, for $i = 1, 2, \ldots, J$. As a result, the classification tree will be build based on rectangular sets S_i produced by recursive partitioning of the data set by one X predictor variable at a time.

For example, let us consider a response variable with five classes and two predictor variables X_1 and X_2. Let us assume the first partition occurs at the value $X_1 = \alpha_1$ such that observations with X_1 values less than or equal to α_1 are assigned to the left branch of the tree, and the observations with X_1 values greater than α_1 are assigned to the right branch of the tree at the root node. That is at $X_1 = \alpha_1$, the gain in impurity measure ΔI at that node should be maximized. Next, the observations with $X_1 \leq \alpha_1$ and $X_2 \leq \beta_1$ are assigned to the first leaf node S_1. In a similar fashion, observations with $X_1 \leq \alpha_1$ and $X_2 > \beta_1$ are assigned to the second leaf node S_2. The recursive partitioning shown in the branches in Figure 11.7.2b follow a similar logic. Figure 11.7.2a shows the complete partition of the X-space into S_1, S_2, S_3, S_4, and S_5.

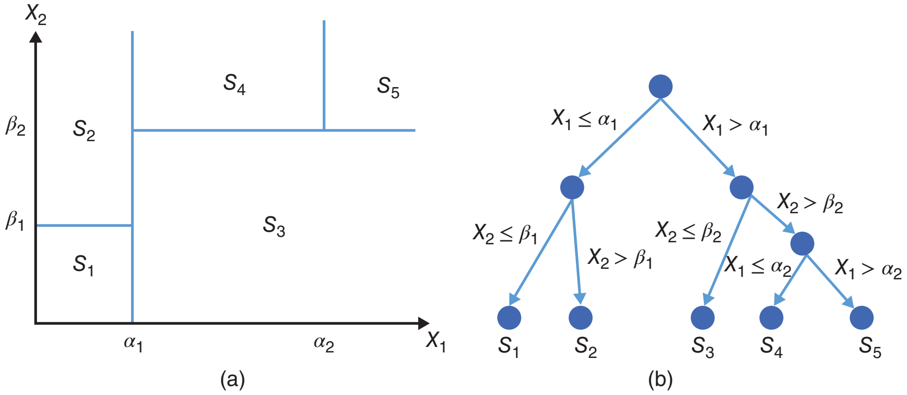


Figure 11.7.2 (a) Recursive partition of the X-space. (b) Corresponding binary tree structure for the partition space in (a).

In classification trees, given a set of observations, the impurity simply measures the proportion of those observations belong to the same class. There are several measures to quantify the impurity, namely, *Gini index*, *entropy/information*, and *classification error*.

The definitions of Gini and entropy measures are given below.

Definition 11.7.1 Suppose the response variable Y has J classes and let p_i be the probability that the observations belong to the ith class of Y, where $i \in \{1, 2, \ldots, J\}$. Then, the Gini impurity measure at the node Q, $Gini(Q)$, is defined as

$$Gini(Q) = 1 - \sum_{i=1}^{J} p_i^2 \tag{11.7.1}$$

The Entropy measure, $I(Q)$ is defined as

$$I(Q) = -\sum_{i=1}^{J} p_i \log_2 p_i \tag{11.7.2}$$

The Gini index based measure for quality of a split due to splitting a parent node into k different nodes is

$$Gini_{split} = \sum_{j=1}^{k} \frac{|n_j|}{|n|} Gini(Q_j) \tag{11.7.3}$$

where n_j is the number of observations that belong to node Q_j after the suggested split and n is the total number of observations in the parent node prior to the suggested split. CART finds the best splitting criteria which minimizes the $Gini_{split}$. In a similar fashion, we can define the gain in entropy information which is commonly discussed in data mining texts (see Kantardzic, 2011).

Example 11.7.1 *To understand the impurity calculation, we consider the data set in Table 11.7.1 with three predictors A, B, C, and response* Y. *As it can be seen in the data, the variables* A, B, *and* Y *are class variables, and the variable* C *is continuous.*

Table 11.7.1 Sample data set with four variables.

A	B	C	Y	A	B	C	Y
1	1	0.2	T	0	2	1.1	F
1	1	0.2	T	0	2	1.0	F
1	2	0.2	T	0	2	1.3	F
1	1	0.2	T	0	2	1.2	F
1	1	0.1	T	0	3	1.5	F
1	1	0.2	T	0	3	1.5	F
1	1	0.1	T	0	3	1.4	F
1	1	0.3	T	0	3	1.5	F
1	2	1.3	F	0	3	1.4	F
0	2	1.1	F	0	3	1.4	T

Table 11.7.2 shows the cross tabulated variables A, B, and C versus the response Y. For example, variable A has 11 "class 0" observations which give rise to 10 "F" responses and 1 "T" response, and it has 9 "class 1" observations which give rise to 1 "F" responses and 8 "T" responses. Based on the Gini index equation in (11.7.1), we calculate the Gini indices for both the classes of this variable. For variable A, Gini index for class 0 is $Gini(A_0) = 1 - \sum_{i=1}^{2} p_i^2 = 1 - (10/11)^2 - (1/11)^2 = 0.165$ and for class 1 is $Gini(A_1) = 1 - \sum_{i=1}^{2} p_i^2 = 1 - (1/9)^2 - (8/9)^2 = 0.198$.

Without loss of generality, the process is repeated for variable B with three classes ("1," "2," and "3"). However, the continues variable C needs to be dichotomized to obtain possible splits, and we decided 0.3 as the cutoff value to split variable C into two classes based on the obvious data separation that occurs at $C = 0.3$ as shown in Figure 11.7.3. Then, we calculate the Gini indices for both the resulting classes as reported in Table 11.7.2.

The quality of the suggested splits is measured using the $Gini_{split}$ index in Equation (11.7.3). The resulting indices due to possible splitting of variables A, B, and C are as follows:

For variable A: $Gini_A = (11/20)0.165 + (9/20)0.198 = 0.180$
For variable B: $Gini_B = (7/20)0 + (7/20)0.245 + (6/20)0.278 = 0.169$
For variable C: $Gini_C = (12/20)0.153 + (8/20)0 = 0.092.$

Table 11.7.2 Gini index computation for variables in Table 11.7.1.

For variable A			
	True Y class		
A	F	T	Gini index
0	10	1	$1-(10/11)^2-(1/11)^2=0.165$
1	1	8	$1-(1/9)^2-(8/9)^2=0.198$
For variable B			
	True Y class		
B	F	T	Gini index
1	0	7	$1-(0/7)^2-(7/7)^2=0.000$
2	6	1	$1-(6/7)^2-(1/7)^2=0.245$
3	5	1	$1-(5/6)^2-(1/6)^2=0.278$
For variable C			
	True Y class		
C	F	T	Gini index
>0.3	11	1	$1-(11/12)^2-(1/12)^2=0.153$
≤ 0.3	0	8	$1-(0/8)^2-(8/8)^2=0.000$

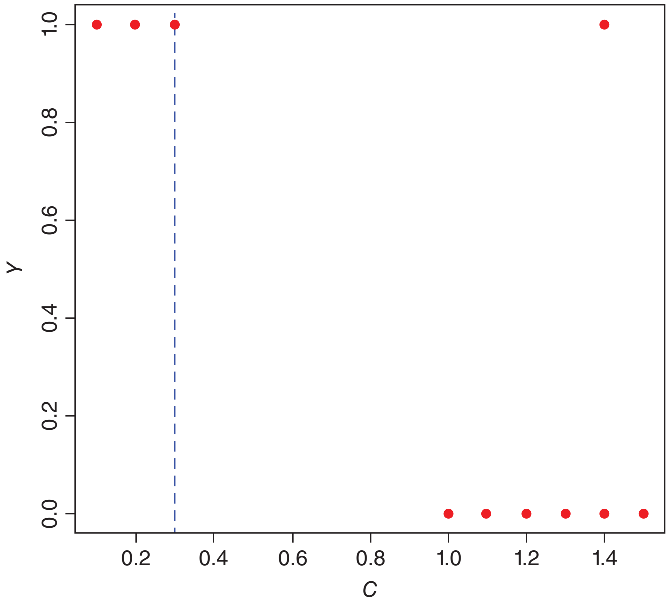


Figure 11.7.3 Graph of variable C versus Y for data in Table 11.7.1.

It is clear from the above calculation that the splitting rule of variable C produces the minimum $Gini_{split}$ value ($Gini_C = 0.092$). Therefore, the tree must split the root node into two branches based on the variable C such that the observations with C values less than or equal to 0.3 (there are 8 observations) lead to left branch, and the rest of the observations lead to the right branch as shown in the Figure 11.7.4.

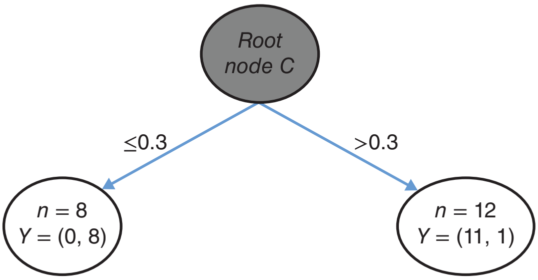


Figure 11.7.4 Initial tree branches based on Gini index.

Example 11.7.2 (Iris Flower Data) *We reconsider the Iris flower data we discussed in Example 11.4.2. First, we use these Iris data to build (or train) a classification tree using 80% of its data and validate the fitted tree using the rest of 20% of data. As mentioned earlier, Table 11.4.3 shows the first few observations of the data set, and the complete data set is available on the website: www.wiley.com/college/gupta/statistics2e for download. We use the first 116 observations to train the model. The "Species" is the response variable with three classes (Setosa, Versicolor, and Virginica) and Sepal Length, Sepal Width, Petal.Length, and Petal Width are the predictors.*

We will handle this data using a computationally powerful "party" package in the R software. The package can handle various types of data including missing, censored, and multivariate data. The "ctree()" procedure in the "party" library is a nonparametric class of tree technique which uses the conditional inference procedures. As this chapter covers only the basic tree concepts, we discarded its advanced theories, and the reader should refer to Hothorn et al. (2006, 2015) for more information. However, due to its usefulness and popularity, we applied the "ctree()" procedure in our example. The R-code for all the outputs is given at the end of this subsection.

The resulting classification tree for Iris data via "ctree" procedure is shown in Figure 11.7.5. "Petal length" is the primary variable that splits the root node that classifies the first leaf node with 42 Setosa when "Petal.Length ≤1.9." When "Petal.Length >1.9," the tree produces a child node (node 3) that splits into two branches based on the variable "Petal.Width," either ≤1.7 or >1.7. Subsequently, if "Petal.Width ≤1.7," the tree produces node 4 that splits into two branches to produce two leaf nodes depending on "Petal.Length" is ≤4.7 or >4.7. Finally, when "Petal.Length >1.9" and "Petal.width >1.7," the tree produces the leaf node 7. Further explanations of the quantities given in the R output can be obtained from the "ctree˙control()" function in R.

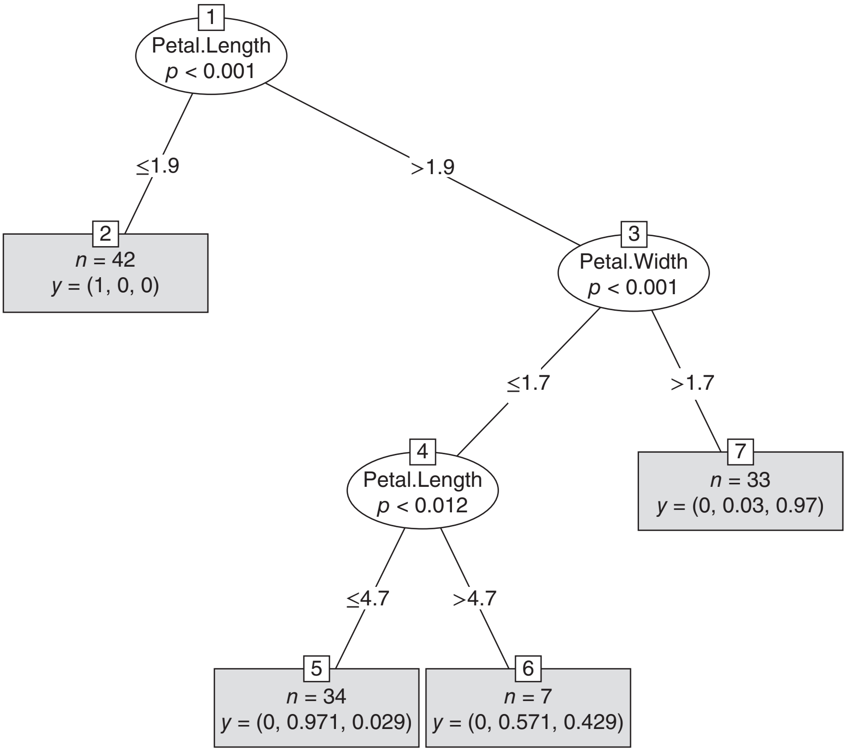


Figure 11.7.5 Classification tree structure for the Iris data. The vector y arranged in (Setosa, Versicolor, Virginica) order.

It is clear from the tree that some nodes such as node 2, node 5, and node 7 provide very strong classifications criterion, but node 6 tends to misclassify between Versicolor and Virginica. In summary, the tree seems to classify flowers well in general, but it would be interesting to see the misclassification rate for this test data set. Table 11.7.3 shows the confusion matrix of training data. It shows that only five observations being misclassified (4 Virginicas as Versicolors and 1 Versicolor as a Virginica) out of 116 observations. Therefore, the resulting overall error is

$$\text{Overall error rate} = \frac{5}{116} = 0.0431$$

Now, using the test data set (observations 117–150 associated with Table 11.4.3) to evaluate the accuracy of the fitted tree, and as shown in the confusion matrix (Table 11.7.4), that tree misclassifies one Virginica (observation # 146 of the data set) flower as a versicolor leading to an overall classification error of

$$\text{Overall error rate} = \frac{1}{34} = 0.0294.$$

Table 11.7.3 Confusion matrix for original training data in Example 11.7.2.

	Predicted class		
True class	Setosa	Versicolor	Virginica
Setosa	42	0	0
versicolor	0	37	1
Virginica	0	4	32

Table 11.7.4 Confusion matrix for randomly selected 34 Iris flowers (observations 117–150 associated with Table 11.4.3).

	Predicted class		
True class	Setosa	Versicolor	Virginica
Setosa	8	0	0
versicolor	0	12	0
Virginica	0	1	13

USING R

The built in R function "ctree()" in "party" library can be used to construct classification trees in R. The *Iris* data explained in Example 11.7.2 is freely available for public use in R. Therefore, one can use the following R code to split data into training and testing data. However, this particular data set is included in Table 11.4.3 (see website: www.wiley.com/college/gupta/statistics2e) so that one can import data directly from a local directory.

```
install.packages("party")
library(party)


#Split data into training and testing data
set.seed(12345)
ind = sample(2, nrow(iris), replace=TRUE, prob=c(0.8, 0.2))
trainData = iris[ind==1,]
testData = iris[ind==2,]
```

The following R code can be used to obtain the classification tree and to obtain the confusion matrices.

```
#Fitting a classification tree with an appropriate “model”
model = Species ∼ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
train.tree = ctree(model, data=trainData)


print(train.tree)
plot(train.tree, type=“simple”)


# Check the prediction for training data
table(trainData$Species, predict(train.tree))


# Predict on test data
testPred = predict(train.tree, newdata = testData)
table(testData$Species, testPred)
```

Regression Trees

In the cases where the outcome variable is at least ordinal, a *regression tree* approach can be used to successively partition observations into subgroups similar to the classification tree approach. In this method, a regression model is fitted to each node to obtain the predicted value at that node. As a result, we obtain a piecewise constant model and that can be used to estimate the outcome of a new observation based on its underline partition. However, the way we measure the impurity is different from the classification tree approach. In the classification tree approach, we defined the impurity of a partition based on those observations that belong to the same class. In regression trees, we use the variance reduction technique that minimizes the total variability results in a split. In other words, we sequentially identify the splitting rules that provide the smallest within-group variances of the outcome variable due to the partitions.

Example 11.7.3 (Stage C Prostate Cancer Data) *This example is related to a data set about 146 Stage C prostate cancer patients reported in Nativ et al. (1988). The response variable of interest is “time to progression,” and the predictor variables are given below.*

- *pgstat: status at last follow-up (1=progressed, 0=censored)*
- *age: age at diagnosis*
- *eet: early endocrine therapy (1=no, 2=yes)*
- *ploidy: diploid/tetraploid/aneuploid DNA pattern*
- *g2: % of cells in G2 phase*
- *grade: tumor grade (1–4)*
- *gleason: Gleason grade (3–10)*

In this example, we exhibit the use of a regression tree approach in estimating "time to progression" in Stage C prostate cancer patients. We use 132 randomly selected observations to train our regression tree and later we will use the rest of the 14 observations to exhibit the accuracy of the fitted model. The first 15 observations of this data set is shown in Table 11.7.5. The complete data set is available on the website: www.wiley.com/ college/ gupta/ statistics2e.

Table 11.7.5 Stage C prostate cancer data (only the first few observations are shown here).

pgtime	pgstat	age	eet	g2	grade	gleason	ploidy
6.1	0	64	2	10.26	2	4	diploid
9.4	0	62	1	NA	3	8	aneuploid
5.2	1	59	2	9.99	3	7	diploid
3.2	1	62	2	3.57	2	4	diploid
4.8	0	69	1	6.14	3	7	diploid
5.8	0	75	2	13.69	2	NA	tetraploid
7.3	0	71	2	NA	3	7	aneuploid
3.7	1	73	2	11.77	3	6	diploid
15.9	0	64	2	27.27	3	7	tetraploid
2.9	1	58	2	14.82	4	8	tetraploid
1.5	1	70	2	10.22	3	8	diploid
14.5	0	67	2	15.66	2	6	tetraploid
4.2	1	66	2	17.79	3	7	tetraploid
1.7	1	74	2	11.11	3	8	diploid
5.0	0	70	2	11.44	2	5	diploid

As shown in the regression tree in Figure 11.7.6, the variable "gleason" is the primary variable that made the initial split. That is, the left branch is formed at the root node for the observations with Gleason grade greater than or equal to 6, and the right branch is formed at root node for the observations with Gleason grade less than 6. Then, the first children nodes split based on the variables "g2" (first left child node) and "age" (first right child node). In similar fashion, we recursively partition the variable space to obtain a tree of depth 7.

After training the regression tree as shown in Figure 11.7.6, we use our test data set with 14 observations to test the prediction ability of the fitted model. If the predictions are accurate, we expect a linear association between the original and predicted values of the response variable "time to progression." However, for our example, the trend is not

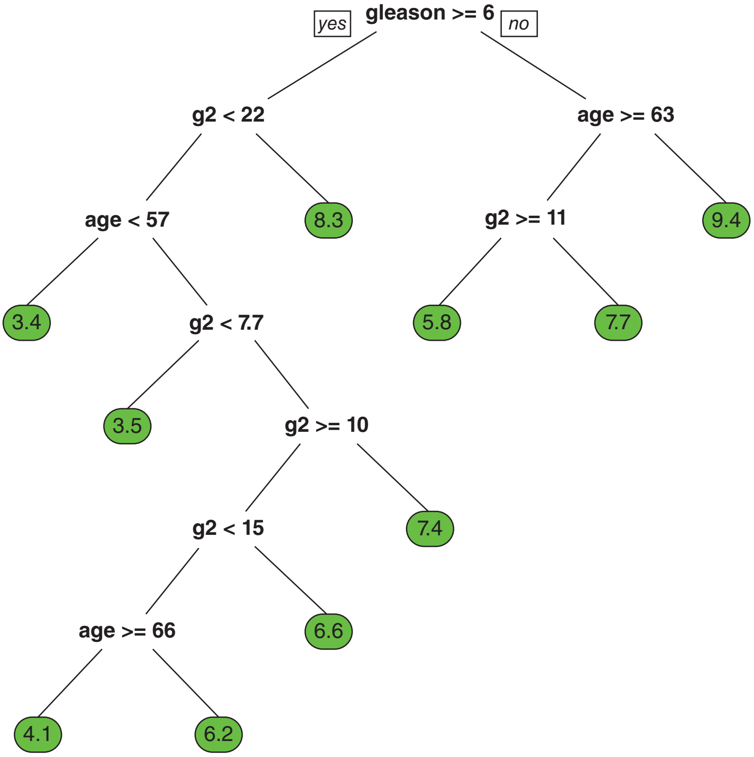


Figure 11.7.6 Regression tree structure for the first 132 observations selected from the Stage C prostate cancer data in Table 11.7.5.

that linear (see Figure 11.7.7) and that indicates the predictive ability of the estimated regression tree is not sufficient. However, the predictive ability of this model can further be developed by techniques such as *bagging, boosting*, and *random forest*.

USING R

The built in R function “rpart()” in “rpart” library can be used to construct regression trees in R. Also, the additional R library “rpart.plot” is also required to obtain a plot of better quality regression tree in R. The *Stage C Prostate Cancer Data* explained in Example 11.7.3 is freely available for public use in R. Therefore, one can use the following R code to spilt the data into training and testing data sets. However, this particular data set is included in Table 11.7.5 so that one can import data directly from a local directory. Note that this data set has missing values, and “rpart” function uses a surrogate approach to estimate the missing values (see Therneau and Atkinson, 1997).

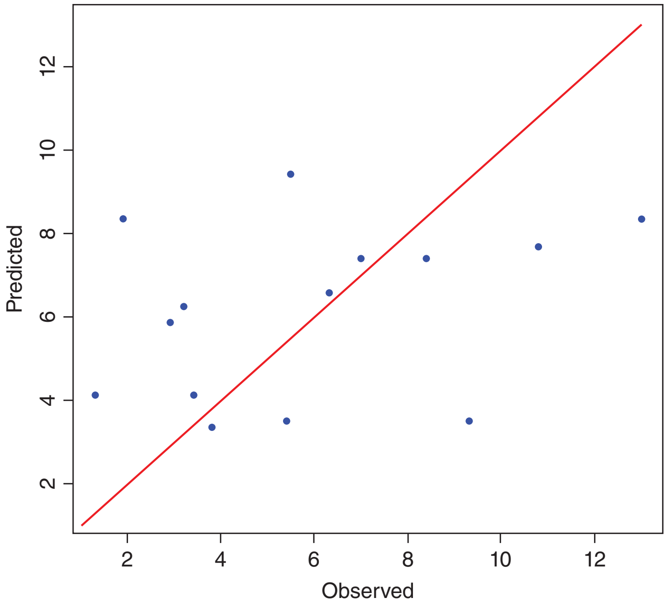


Figure 11.7.7 Observed and predicted “time to progression” values for the test data (the last 14 observations selected from the Stage C prostate Cancer data in Table 11.7.5).

```
install.packages('rpart'); install.packages('rpart.plot')
library(rpart); library(rpart.plot)

#Split data into training and testing data
set.seed(123)
ind = sample(2, nrow(stagec), replace=TRUE, prob=c(0.9, 0.1))
trainData = stagec[ind==1,]
testData = stagec[ind==2,]
```

The following R code can be used to obtain the regression tree and required predictions.

```
#Fitting a regression tree with an appropriate 'model'
model = pgtime ~ age + eet + g2 + grade + gleason + ploidy
reg.tree = rpart(model, data = trainData)

#Plot the regression tree
par(xpd=TRUE)
prp(reg.tree, faclen = 0, cex = 1, box.col = 'green')

#Predictions of test data plot against true responses
stagec.pred = predict(reg.tree, newdata=testData)
xlim = range(testData$pgtime)
plot(stagec.pred ~ pgtime, data=testData, xlab='Observed', ylab='Predicted',
ylim=xlim, xlim=xlim, col=4, cex=1.5, pch=20)
lines(c(1,13), c(1,13), col=2, lwd=2)
```

11.7.2 Further Reading

As we mentioned earlier in this chapter, due to some space limitations, our goal in this chapter is just to introduce some basic concepts in data mining and discuss classification and regression trees procedures. However, we recommend the reader to read some of the references on this topic. The classic book on classification and regression trees is by Breiman et al. (1984), and chapter 1 on trees in Ripley (1996). Some other references are Reid (1982), Mitchell (1997), Han et al. (2011), and Kantardzic (2011).

11.8 CASE STUDIES

Case Study 1 (*Heart Disease*)[1] Data set contains 14 selected attributes from 303 heart patients from the Cleveland database. Naturally, this data contain both quantitative and qualitative attributes, but for the analysis purpose, the qualitative attributes have been coded using numerical scales. The attribute "target" indicates the angiographic disease status (0 = absence, 1 = presence), and the main purpose of this case study is to build a classification tree that can be used to predict the "target." Attributes information are as follows:

1. age
2. sex (0 = female, 1 = male)
3. cp = chest pain type (4 values)
4. trestbps = resting blood pressure
5. chol = serum cholestoral in mg/dl
6. fbs = fasting blood sugar >120 mg/dl
7. restecg = resting electrocardiographic results (values 0, 1, 2)
8. thalach = maximum heart rate achieved
9. exang = exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. slope = the slope of the peak exercise ST segment
12. ca = number of major vessels (0–3) colored by fluoroscopy
13. thal (3 = normal; 6 = fixed defect; 7 = reversible defect)

The data reported for this Case Study are available under *case study 11.8.1* on the book website: www.wiley.com/college/gupta/statistics2e.

(a) Randomly split this data set into training and testing sets using 70:30 ratio.
(b) Construct a classification tree using all 13 attributes to predict the "target."
(c) Identify the significant attributes in predicting the variable "target."
(d) Discuss the quality of the fitted model in part (b) using its confusion matrix.
(e) Repeat parts (b)–(d) by varying the data ratio to 60:40 in training and testing data sets in part (a).
(f) Explain any interesting finding from part (e).

[1] **Source**: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. Data available at https://www.kaggle.com/ronitf/heart-disease-uci and more information can be found at https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Case Study 2 (*King County House Data*)[2] This study focuses on building a regression tree to predict house prices in King county, Washington. Data show the house pricing in 2014–2015 time period and include 21 unique attributes. An analyst is interested in building a regression tree to predict house prices ("price") using the following 16 attributes shown in data sets *Case Study 11.8.2a* and *Case Study 11.8.2b* on the book website: www.wiley.com/college/gupta/statistics2e.

1. bedrooms = number of bedrooms
2. bathrooms = number of bedrooms
3. sqft.living = total area of the living area
4. sqft.lot = total area of the lot
5. floors = number of floors
6. waterfront = adjacent to a water front (0 = no, 1 = yes)
7. view = nature of the view (ranging 0–4)
8. condition = nature of the condition (ranging 1–5)
9. grade = grade of the house (ranging 1–13)
10. sqft.above = total living area above the basement
11. sqft.basement = total area of the basement
12. yr.built = year built
13. yr.renovated = year renovated
14. zipcode = zip code
15. lat = latitude
16. long = longitude

Change the variable "yr.renovated" to a dichotomous variable that reflects whether that house is renovated or not after building and "sqft.basement" to a dichotomous variable that reflects whether there is a basement or not in that house.

(a) Construct a regression tree using all 16 attributes to predict the "price."
(b) Identify the significant attributes in predicting the variable "price."
(c) Use the fitted regression tree in part(a) to predict house prices shown in data set *Case Study 11.8.2b.*
(d) Evaluate the quality of the predictions obtained in part (c) using a suitable method.

11.9 USING JMP

This section is not included in the book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. What is data mining and what is big data?
2. What are the three key points one should consider in the process of data reduction and preparing for data mining? Explain.

[2] **Source**: https://www.kaggle.com/andres111mejia/multivariate-regression/log

3. What are the two main ways one can visualize data? Explain.

4. Consider a part of the famous “Titanic” data set[3] listed in “Review Problem 4” (see website: www.wiley.com/college/gupta/statistics2e). It includes the following variables:

   Passenger ID: Identification number for data analysis purpose.
   Survival: 0 = No, 1 = Yes.
   Pclass: Ticket class, 1 = 1st, 2 = 2nd, 3 = 3rd.
   Sex: Male or female.
   Age: Age in years.
   SibSp: # of siblings and spouses aboard the Titanic.
   Parch: # of parents and children aboard the Titanic.
   Fare: Passenger fare.
   Embarked: Port of Embarkation, C = Cherbourg, Q = Queenstown, S = Southampton.

   (a) List basic summary statistics for each variable.
   (b) Use appropriate tools to graph each variable.

5. Address the precautions one can take to deal with the missing values exist in the data given in Review Problem 4.

6. Conduct a thorough outlier analysis for the appropriate variables given in data in Review Problem 4.

7. Consider the data given in Review Problem 4.

   (a) Obtain a 2D scatter plot of variables “Fare” versus “Age” and identify the nature of survival considering “Survival” as the third variable.
   (b) Obtain a 3D scatter by plotting variables “Age,” “Fare,” and “Survival” against each other.
   (c) Comment on your findings in parts (a) and (b).

8. Consider the data given in Review Problem 4.

   (a) Obtain a side-by-side boxplots for “Fare” for “Pclass” and comment on your graphs.
   (b) Obtain an aesthetic scatter plot for “Age,” “Fare,” “Sex,” and “Survival” and comment on your graph.

9. Consider the data given in Review Problem 4.

   (a) Obtain a contingency table to investigate the relationship between “Sex” and “Survival.” Comment on the relationship(s).
   (b) Obtain a contingency table to investigate the relationship between “Pclass” and “Survival.” Comment on the relationship(s). (*Hint*: you may perform the appropriate conditional probability calculations.)

10. Consider the data given in Review Problem 4.

    (a) Obtain a contingency table to investigate the relationship between “SibSp” and “Survival.” Comment on the relationship(s).
    (b) Obtain a contingency table to investigate the relationship between “Parch” and “Survival.” Comment on the relationship(s).
    (c) Obtain a contingency table to investigate the relationship between “Embarked” and “Survival.” Comment on the relationship(s). (*Hint*: you may perform the appropriate conditional probability calculations.)

[3] **Source:** https://www.kaggle.com/c/titanic/data.

11. Use the following data set to obtain the confusion matrix and report both class 0 and class 1 misclassification rates.

Age	Status	Predicted
38	0	0
43	1	1
43	1	1
39	0	1
43	1	1
36	0	0
31	0	0
39	0	1
39	1	1
39	1	1
43	1	1
42	1	1
35	0	0
44	1	1
49	0	0

12. Consider the data given in Review Problem 11.
    (a) Use the logistic regression concepts discussed in Section 11.6 (see Section 16.8 for more details) to predict the outcome (i.e., "Status") using the variable "Age."
    (b) Use a cutoff value of 0.5 to obtain the confusion matrix for your predictions in part (a), and report both class 0 and class 1 classification errors.

13. Investigate the relationship between cutoff values (0–1) and misclassification error rates for the predicted probabilities using the logistic model used in Review Problem 12.
    (a) Plot both class 0 and class 1 classification errors against the range of cutoff values (from 0 through 1).
    (b) What would be a reasonable cutoff value that minimizes both class 0 and class 1 classification errors.

14. Consider the data given in Review Problem 4.
    (a) Use the logistic regression concepts discussed in Section 11.6 (see Section 16.8 for more details) to predict the outcome (i.e., "Survival") using the variables "Parch," "Fare," "Embarked." Please disregard the observations with missing values.

(b) Use a cutoff value of 0.5 to obtain the confusion matrix for your predictions in part (a) and report both class 0 and class 1 classification errors.

15. The data listed in "Review Problem 15" (see website: www.wiley.com/college/gupta/statistics2e) contain data from 252 people reported in "Generalized body composition prediction equation for men using simple measurement techniques" by Penrose et al. (1985). The variables listed in the original data include: "Percent of body fat index (BFI)," "Age (years)," "Weight (lbs)," "Height (inches)," "Neck circumference (cm)," "Chest circumference (cm)," "Abdomen circumference (cm)," "Hip circumference (cm)," "Thigh circumference (cm)," "Knee circumference (cm)," "Ankle circumference (cm)," "Biceps (extended) circumference (cm)," "Forearm circumference (cm)," and "Wrist circumference (cm)." For learning propose, we dichotomize the percent of body fat index into a new variable name "Index."
    (a) Construct a classification tree using all the variables except "Percent of body fat index" to predict the "Index."
    (b) Discuss the quality of the fitted model in part (a) using the confusion matrix.
    (c) Construct a regression tree using all the variables except "Index" to predict "Percent of body fat index."
    (d) Evaluate the quality of the prediction using a suitable method.

16. Split the data in Review Problem 15 into training and testing sets, using a 80:20 ratio. That is, use observations 1–202 as training data to build a classification tree using all the variables except "Percent of body fat index" to predict the "Index," and use observations 203–252 as testing data to validate your classification tree. Report the overall classification error.

17. Split the data in Review Problem 15 into training and testing sets, using a 80:20 ratio. That is, use observations 1–202 as training data to build a regression tree using all the variables except "Index" to predict the "Percent of body fat index," and use observations 203–252 as testing data to validate your regression tree. Report your regression tree for training data.
    (a) Plot predicted values of "Percent of body fat index" versus true "Percent of body fat index" in testing data. Describe the trend.
    (b) Report the correlation coefficient between the predicted values of "Percent of body fat index" versus true "Percent of body fat index," and explain the predictability of your model.

18. A professor in a certain college is interested in predicting final class grades of his students using some data prior to the exam. He recorded the following data that include students quiz grades (Q1–Q5), homework grades (H1–H7), mid-term exam scores (T1 & T2), final exam score (Final), and final letter grade (Grade).

Q1	Q2	Q3	Q4	Q5	H1	H2	H3	H4	H5	T1	T2	Final	Grade
10.0	6.0	10.0	10.0	10.0	9.6	10.0	10.0	10.0	10.0	88	91	95	A
10.0	10.0	10.0	10.0	10.0	9.8	9.0	8.9	9.9	9.6	102	100	100	A
10.0	8.0	7.0	7.0	7.0	9.4	10.0	10.0	9.7	9.3	40	47	71	D
10.0	10.0	10.0	10.0	10.0	9.9	9.7	9.4	9.8	9.8	46	93	90	B
10.0	10.0	10.0	10.0	10.0	9.7	9.7	9.7	10.0	9.5	77	97	77	B
7.0	10.0	10.0	10.0	0.0	0.0	7.8	7.9	9.5	4.0	60	90	87	C
10.0	10.0	10.0	7.0	10.0	9.5	9.8	10.0	9.4	10.0	83	73	86	B
10.0	10.0	10.0	0.0	10.0	0.0	7.7	8.9	0.0	0.0	103	70	95	B
10.0	7.0	7.0	10.0	10.0	8.6	9.0	8.4	9.3	8.5	81	74	82	B
10.0	10.0	6.0	10.0	0.0	8.4	9.0	10.0	9.8	10.0	67	76	60	C
7.0	6.0	7.0	10.0	10.0	9.6	9.8	9.9	9.8	10.0	62	65	55	C
10.0	5.0	10.0	0.0	5.0	9.4	9.3	10.0	9.2	10.0	48	47	52	F
10.0	9.0	7.0	10.0	10.0	5.6	8.7	0.0	0.0	0.0	75	70	85	C
10.0	10.0	10.0	0.0	10.0	8.1	10.0	7.5	0.0	0.0	99	87	54	C
10.0	6.0	10.0	10.0	10.0	8.0	13.5	14.0	13.0	11.5	100	77	89	A
10.0	6.0	10.0	7.0	0.0	8.0	14.0	16.0	13.8	11.0	79	60	72	C
10.0	10.0	10.0	4.0	10.0	8.5	13.5	12.0	8.0	0.0	115	88	82	B
10.0	0.0	10.0	7.0	10.0	4.0	7.0	15.0	0.0	5.0	94	42	52	D
0.0	10.0	10.0	10.0	10.0	0.0	0.0	0.0	0.0	0.0	100	38	49	D
7.0	6.0	10.0	7.0	0.0	5.0	8.0	0.0	2.5	0.0	72	0	0	F
7.0	4.0	10.0	7.0	0.0	7.0	13.0	11.0	14.0	0.0	48	27	54	F
7.0	5.0	10.0	7.0	8.0	4.0	14.0	16.0	13.0	0.0	87	46	70	D
10.0	10.0	10.0	10.0	10.0	1.0	14.0	15.0	14.0	12.0	115	95	98	A
10.0	10.0	10.0	0.0	10.0	1.0	7.0	0.0	10.0	0.0	87	48	82	C
10.0	9.0	4.0	9.0	8.0	4.0	9.3	9.8	0.0	0.0	92	74	100	A
10.0	10.0	8.0	10.0	0.0	10.0	10.0	0.0	9.5	0.0	72	73	63	C
10.0	0.0	6.0	0.0	4.0	0.0	9.8	9.0	0.0	0.0	78	40	74	D
10.0	10.0	8.0	0.0	0.0	10.0	9.8	10.0	10.0	10.0	85	72	48	D
7.0	10.0	8.0	9.0	7.0	9.0	10.0	10.0	10.0	10.0	84	70	75	B
7.0	8.0	10.0	10.0	10.0	10.0	10.0	0.0	8.5	0.0	81	50	61	C
10.0	4.0	6.0	8.0	7.0	10.0	10.0	10.0	10.0	10.0	66	74	75	C
10.0	10.0	6.0	0.0	7.0	8.0	8.0	9.3	9.0	9.0	88	35	77	C
7.0	10.0	6.0	10.0	7.0	6.0	0.0	0.0	6.0	0.0	77	49	77	C
10.0	10.0	10.0	9.0	10.0	9.5	8.0	0.0	8.0	4.8	86	80	90	B
10.0	4.0	8.0	0.0	0.0	10.0	10.0	10.0	10.0	2.0	81	74	77	B
10.0	10.0	6.0	10.0	7.0	10.0	10.0	10.0	10.0	10.0	92	68	82	B
10.0	10.0	8.0	10.0	10.0	10.0	9.5	9.5	10.0	9.0	72	72	90	B
0.0	0.0	8.0	10.0	7.0	0.0	6.0	8.5	0.0	3.5	89	57	89	C
10.0	10.0	8.0	0.0	0.0	10.0	10.0	10.0	10.0	9.3	98	69	87	B
7.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	86	85	93	A
10.0	10.0	8.0	0.0	7.0	10.0	9.8	10.0	10.0	0.0	83	66	75	B

Q1	Q2	Q3	Q4	Q5	H1	H2	H3	H4	H5	T1	T2	Final	Grade
10.0	10.0	8.0	0.0	10.0	10.0	10.0	10.0	0.0	6.5	92	75	98	B
10.0	10.0	10.0	10.0	7.0	9.0	9.5	10.0	0.0	10.0	87	65	77	C
10.0	10.0	8.0	10.0	10.0	10.0	10.0	10.0	9.8	9.5	100	99	97	A
10.0	10.0	8.0	0.0	10.0	9.0	10.0	9.0	10.0	7.0	75	54	79	B
10.0	10.0	6.0	10.0	8.0	6.5	10.0	10.0	10.0	7.5	80	86	95	A

(a) Construct a classification tree using the students quiz grades (Q1–Q5), homework grades (H1–H5), mid-term exam scores (T1 & T2) to predict the students final letter grade (Grade).
(b) Discuss the quality of the fitted model in part (a) using the confusion matrix.
(c) Construct a regression tree using the students quiz grades (Q1–Q5), homework grades (H1–H5), and mid-term exam scores (T1 & T2) to predict the students final exam score (Final).
(d) Evaluate the quality of the prediction of model in part (c) using a suitable method.

19. Use the following testing data to predict outcomes of the classification tree derived in Review Problem 18. Discuss the accuracy of the fitted model, using an appropriate measure.

Q1	Q2	Q3	Q4	Q5	H1	H2	H3	H4	H5	T1	T2	Final	Grade
10.0	10.0	10.0	8.0	8.0	10.0	12.0	9.5	10.0	9.5	75	69	76	C
10.0	9.0	10.0	10.0	4.0	9.0	12.0	0.0	0.0	0.0	86	60	59	C
10.0	10.0	10.0	10.0	8.0	10.0	12.0	9.5	8.5	9.5	86	73	85	B
7.0	9.0	8.0	10.0	10.0	10.0	7.0	0.0	8.5	8.0	58	50	83	C
10.0	0.0	0.0	0.0	8.0	0.0	10.5	0.0	0.0	0.0	86	30	59	D
7.0	10.0	7.0	10.0	8.0	0.0	6.5	5.5	0.0	0.0	72	87	100	B
7.0	10.0	10.0	10.0	8.0	10.0	12.0	10.0	10.0	10.0	98	87	93	A
0.0	9.0	8.0	9.0	0.0	9.5	12.0	7.5	6.5	10.0	60	46	74	C
10.0	9.0	4.0	6.0	0.0	6.0	0.0	10.0	10.0	9.5	62	45	47	D
10.0	10.0	4.0	10.0	6.0	7.5	8.0	5.0	7.0	0.0	80	75	84	C
10.0	10.0	8.0	10.0	8.0	10.0	12.0	10.0	10.0	10.0	90	55	82	B
10.0	9.0	9.0	10.0	6.0	9.0	11.0	0.0	0.0	7.5	74	40	54	D

20. Use the testing data in Review Problem 18 to predict outcomes of the regression tree in Review Problem 19. Discuss the accuracy of the fitted model using an appropriate measure.

Chapter 12

CLUSTER ANALYSIS

The focus of this chapter is the discussion of basic clustering techniques.

Topics Covered

- Basic concepts of clustering
- Similarity measures
- Hierarchical clustering methods
- Ward's hierarchical clustering method
- Nonhierarchical clustering methods
- K-means method
- Density based clustering methods
- Model based clustering methods

Learning Outcomes

After studying this chapter, the reader will be able to

- Discuss the various types of similarity measures and clustering techniques.
- Distinguish between hierarchical, non-hierarchical and other types of clustering techniques.
- Perform cluster analysis for various types of data.
- Compare various clustering techniques.
- Summarize and interpret the cluster results.
- Use software packages MINITAB, R, and JMP to perform cluster analysis.

12.1 INTRODUCTION

Grouping objects into one or more groups so that the objects within each assigned group are more homogeneous than otherwise is called clustering. Cluster analysis helps discover

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

"natural" groupings of objects on the basis of similarities between the objects. Unlike classification methods, groups and the number of groups are unknown prior to clustering of data. Cluster analysis is an exploratory technique with no assumptions of group structure or number of groups and is often quite helpful to investigate the complex nature of data structures. Analysts could interpret and validate cluster analysis results based on their understanding of the data. Clustering can be achieved by various algorithms that differ in their notion of what constitutes a cluster and how to efficiently find them. Three major types of clustering algorithms are available in the literature, namely, hierarchical, nonhierarchical, and model-based methods. Applications can be found in a variety of areas such as data mining, pattern recognition, sports, medicine, and bioinformatics (see Johnson and Wichern, 2007; Abonyi and Feil, 2007; Han et al., 2011 for more details).

The following questions in respective disciplines can be answered using appropriate cluster analysis techniques.

***Marketing*:** What type of customers should be targeted for a new product? To answer this question a market analyst may conduct a cluster analysis of customer demographics, income, shopping styles, dining patterns, attitudes, and marital status. This would result in establishing groups of customers that have similar needs and behaviors.

***Bioinformatics*:** What type of disease will a new patient be diagnosed with in the future? To answer this question, a researcher may need to collect information about patient's organs, genes, and proteins along with their basic demographics. Then the cluster analysis can be used to group people with similar symptoms, which may reflect some common genes, proteins, and/or demographic patterns. Characteristics of these homogeneous groups of people can possibly be used to estimate the likelihood of a new patient being categorized into one or more groups.

***Sports*:** An aim may be to select the best set of players in various team sports. An analyst can collect players' career data and cluster them with similar statistics. Then it would be possible to correlate teams' winning percentages, with proportions of clustered groups in each team. These relations would shed light on player selection criteria in future games.

We will discuss how one can conduct such cluster analysis for various types of data, using some of the major clustering methods in the subsequent sections. In the next section, we will introduce basic mathematical tools and conceptual background needed to proceed with cluster analysis.

12.2 SIMILARITY MEASURES

The objects which appear to be close together, naturally exhibit similar properties than the objects that appear farther apart. Therefore, it is important to measure the closeness of objects when clustering. The distance between two objects may be used to measure their closeness, and statistical measures such as correlation may be used to determine closeness of variables. In general, similarity measures are employed to measure closeness of objects or variables with respect to certain criteria:

The *Euclidean distance* between two one-dimensional objects located at x and y is written as

$$D_{x,y} = |x - y| \tag{12.2.1}$$

The Euclidian distance between two two-dimensional objects located as $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ is defined as given below:

$$D_{\mathbf{x},\mathbf{y}} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \tag{12.2.2}$$

The general form of the Euclidean distance for two p-dimensional objects located at $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ is

$$D_{\mathbf{x},\mathbf{y}} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \tag{12.2.3}$$

Euclidean distance is a special case of a more general measure known as *Minkowski distance.* For two p-dimensional objects located at $\mathbf{x}$ and $\mathbf{y}$, the Minkowski distance is

$$D_{\mathbf{x},\mathbf{y}} = \left[\sum_{i=1}^{p} |x_i - y_i|^w\right]^{1/w} \tag{12.2.4}$$

The Minkowski distance becomes the *city-block distance* (or *Manhattan distance*) when $w = 1$, and becomes the two-dimensional Euclidean distance when $w = 2$.

The cosine correlation coefficient is another useful similarity measure. To define the cosine correlation coefficient we first consider the Euclidean *dot product* formula between two objects located at $\mathbf{x}$ and $\mathbf{y}$. That is,

$$\mathbf{x}.\mathbf{y} = |\mathbf{x}||\mathbf{y}|\cos(\theta) \tag{12.2.5}$$

where θ is the angle between the two vectors $\mathbf{x}$ and $\mathbf{y}$ and $|\mathbf{x}| = \sqrt{\sum_{i=1}^{p} x_i^2}$, $|\mathbf{y}| = \sqrt{\sum_{i=1}^{p} y_i^2}$ are the lengths of the vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. Then we have from (12.2.5) $\cos(\theta)$, the cosine correlation coefficient, which may be written as

$$\cos(\theta) = \frac{\mathbf{x}.\mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2}\sqrt{\sum_{i=1}^{p} y_i^2}} \tag{12.2.6}$$

This similarity measure ranges from -1 to 1 indicating perfect opposites to perfect similarity, and a value near zero indicates total dissimilarity. If we use the standardized vectors in (12.2.6), it would provide a measure equivalent to the Pearson correlation coefficient.

Given a set of objects, we can summarize all the pairwise distances in a *distance matrix.* Also, we can visualize a distance matrix using many different methods to identify vital distances between objects. Now we exhibit one graphical display using the network graphing method in the following example.

Example 12.2.1 (Crime data) *Consider a set of crime data reported*[1] *in six different US states that includes attributes "Murder rate," "Assault rate," "Rape rate," and "Urban population." Table 12.2.1 shows crime rates per 100,000 residents and the variable, "Urban Pop" represents the percent of the population living in urban areas in the given state.*

Table 12.2.1 Reported crime data.

State	Murder rate	Assault rate	Rape rate	Urban pop
A	15.4	335	31.9	80
B	13.0	337	16.1	45
C	9.0	276	40.6	91
D	11.3	300	27.8	67
E	8.1	294	31.0	80
F	11.4	285	32.1	70

[1] McNeil (1977)

First we calculate the Euclidean distance $D_{A,B}$ between states A and B for all four crime characteristics. From (12.2.3),

$$D_{A,B} = \sqrt{(15.4 - 13.0)^2 + (335 - 337)^2 + (31.9 - 16.1)^2 + (80 - 45)^2} = 38.53$$

Then the city-block distance (with $w = 1$ in (12.2.4)) between states A and B are

$$D_{A,B} = |15.4 - 13.0| + |335 - 337| + |31.9 - 16.1| + |80 - 45| = 55.2$$

Then, we use (12.2.6) to calculate the cosine correlation coefficient between states A and B as follow:,

$$\begin{aligned}\cos(\theta) &= \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2}\sqrt{\sum_{i=1}^{p} y_i^2}}\\ &= \frac{15.4 \times 13.0 + 335 \times 337 + 31.9 \times 16.1 + 80 \times 45}{\sqrt{15.4^2 + 335^2 + 31.9^2 + 80^2}\quad\sqrt{13^2 + 337^2 + 16.1^2 + 45^2}} = 0.994\end{aligned}$$

The resulting cosine correlation coefficient estimate (= 0.994) indicates a higher similarity between states A and B. However, both the distance measures do not provide such information directly, though those distances can be used for comparison purpose as we illustrate below.

We use the R software to calculate all the pairwise Euclidean distances and to obtain a network graph to compare the resulting distances.

USING R

We may use the built-in R function ‘dist()’ to find all pairwise Euclidean distances and plot resulting distances in a network graph using ‘qgraph()’. The following R code can be used to obtain both the distance matrix and the network graph.

```
library(qgraph)
#Data preparation
Murder = c(15.4,13,9,11.3,8.1,11.4)
Assault = c(335,337,276,300,294,285)
Rape = c(31.9,16.1,40.6,27.8,31,32.1)
Urban = c(80,45,91,67,80,70)
crime = cbind(Murder, Assault, Rape, Urban)
row.names(crime) = c(“A”,“B”,“C”,“D”,“E”,“F”)

#To get the distance matrix
distance = dist(crime, method = “euclidean”)
distance

#To get the network graph
Inv.dist = 1/distance #input reciprocal distances
qgraph(Inv.dist, layout=“spring”, vsize=5)
```

Table 12.2.2 shows the resulting Euclidean distance matrix between states with regards to the reported crime rates. The resulting network graph is shown in Figure 12.2.1. The length between any two vertices on the graph indicates the distance between those two objects and the width and the color intensity of the edges indicate closeness of the objects. That is, thickens and the color intensity are inversely proportional to the actual distance between true objects. For the data in Example 12.2.1, the states ‘E’ and ‘F’ provides

Table 12.2.2 Euclidean distances between six US states.

	A	*B*	*C*	*D*	*E*
B	38.5				
C	61.0	80.3			
D	37.8	44.6	36.3		
E	41.7	57.6	23.2	15.0	
F	51.1	59.9	24.5	15.9	13.9

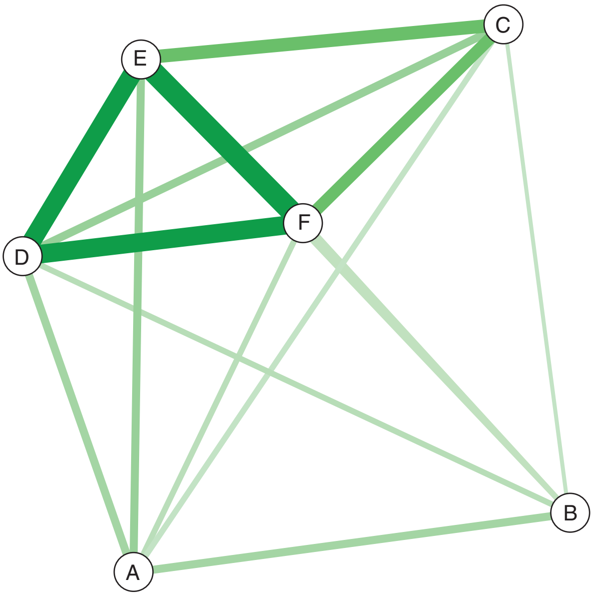


Figure 12.2.1 Network graph for distance matrix in Table 12.2.2.

the minimum pairwise distance among all the states, while ‘E and D’ and ‘F and D’ have the next minimum distances, respectively. All this information is again apparent in Figure 12.2.1, and therefore, one can visualize which states tend to clump together with regards to crime distances.

The distance calculation becomes challenging when the data are nonnumeric. In real life, the data can be a combination of many forms, and their attributes may represent different characteristics. For instance, a data vector can represent both quantitative and qualitative attributes. For example, consider a set of patient data consisting of five-dimensional characteristics such as gender, age, blood type, systolic blood pressure, and glucose level. Attributes such as age, systolic blood, pressure, and glucose level are quantitative, but gender and blood type are qualitative in nature. The following example is used to exhibit the distance and other related calculations for such data.

Example 12.2.2 (A clinical trials study) *Consider the following measurements taken from two male and two female patients in a clinical trials study.*

Table 12.2.3 A clinical trials study data.

Patient ID	Gender	Age (yr)	Blood type	Systolic blood pressure (mm Hg)	Glucose level (mg/dl)
001	F	57	A	125	101
002	M	60	B	130	112
003	M	65	A	120	98
004	F	58	B	128	107

We cannot perform vector calculations on row vectors of the data in Table 12.2.3 due to the qualitative nature of some attributes namely gender and blood type. Therefore, it is required to change the structure of these data to a common numerical structure. Thus, we dichotomize each variable into a binary variable as follows:

$$\text{Let } X_{Gen} = \begin{cases} 1 & \text{if M} \\ 0 & \text{if F} \end{cases}, X_{Age} = \begin{cases} 1 & \text{if Age} \geq 60 \\ 0 & \text{if Age} < 60 \end{cases}, X_{BT} = \begin{cases} 1 & \text{if Blood type is A} \\ 0 & \text{Otherwise} \end{cases},$$

$$X_{BP} = \begin{cases} 1 & \text{if Blood pressure} \geq 125 \\ 0 & \text{if Blood pressure} < 125 \end{cases}, X_{GL} = \begin{cases} 1 & \text{if Glucose level} \geq 100 \\ 0 & \text{if Glucose level} < 100 \end{cases}$$

The resulting data after above coding can be summarized as follows in Table 12.2.4.

Table 12.2.4 Dichotomous attributes of the clinical trials study data.

Patient ID	X_{Gen}	X_{Age}	X_{BT}	X_{BP}	X_{GL}
001	0	0	1	1	1
002	1	1	0	1	1
003	1	1	1	0	0
004	0	0	0	1	1

Now the similarity between items can be calculated after constructing a 2 × 2 contingency table for each pair of items. Let a = number of 1–1 pairs, b = number of 1–0 pairs, c = number of 0–1 pairs, and d = number of 0–0 pairs. Then, the entries of the contingency table are calculated so that the resulting contingency Table 12.2.5 for the patients 001 and 002, is as given below:

Table 12.2.5 The contingency table for patients 001 and 002 based on data from Table 12.2.4.

		Patient 002	
		1	0
Patient	1	$a = 2$	$b = 1$
001	0	$c = 2$	$d = 0$

For patients 001 and 002, there are only two matching pairs $(a + d = 2)$ among the five considered characteristics. That provides $2/5 = 0.40$ matching rate, which can be used for measuring similarity between them. However, we will now define a few additional *similarity coefficients* for clustering items and further discuss this example.

12.2.1 Common Similarity Coefficients

1. The *simple matching coefficient* (SMC) which equally weights both 1–1 and 0–0 matches
$$SMC = \frac{a+d}{a+b+c+d}$$
2. The *Russel and Rao coefficient* (RRC) which excludes 0–0 matches from the numerator
$$RRC = \frac{a}{a+b+c+d}$$
3. The *Jaccard coefficient* (JC) which completely excludes 0–0 matches
$$JC = \frac{a}{a+b+c}$$
4. The *Match-Mismatch coefficient* (MMC) which completely excludes 0–0 matches
$$MMC = \frac{a}{b+c}$$

The Jaccard coefficient for patients 001 and 002 becomes equal to $\frac{a}{a+b+c} = \frac{2}{2+1+2} = 0.40$. That indicates 40% Jaccard coefficient-based matching rate between patients 001 and 002. In a similar fashion, we calculate the Jaccard coefficients for all pairs of patients and summarize them in Table 12.2.6

Table 12.2.6 The Jaccard coefficients for the pairs of patients in Table 12.2.4.

		Patient		
		001	002	003
Patient	002	0.40		
	003	0.20	0.40	
	004	0.67	0.50	0.00

The calculated Jaccard coefficients indicate that patients 001 and 004 are more homogeneous, and the patients 003 and 004 are more heterogeneous compared to the rest.

We can also calculate the Euclidean distance between items using dichotomous coded variables. The squared Euclidean distance $\sum_{i=1}^{p}(x_i - y_i)^2$ measures the number of mismatches. For instance, the number of mismatches for patients 001 and 002 shown in Table 12.2.4 is $\sum_{i=1}^{5}(x_i - y_i)^2 = (0-1)^2 + (0-1)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2 = 3$. Note that this distance measure ignores 1–1 and 0–0 matches and assumes those two cases are equally unimportant.

In some applications, we may have to group the variables, not the objects. In such cases, we calculate Pearson and Spearmen correlation coefficients depending on the data type. For bivariate or dichotomous variables, strength of association can be calculated using the *Pearson product moment correlation* r, where r

$$r = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} \tag{12.2.7}$$

where $a =$ number of $1-1$ pairs, $b =$number of $1-0$ pairs, $c =$ number of $0-1$ pairs, and $d =$number of $0-0$ pairs.

We will illustrate the use of the Pearson product moment correlation calculation and interpretation in Example 12.2.3.

Example 12.2.3 (Clinical trials study data in Example 12.2.2) *Reconsider the dichotomous clinical trials study data set in Table 12.2.4, and calculate the strength of association between variables gender and age, using the Pearson product moment correlation formula, given in (12.2.7).*

Solution: The corresponding contingency table for gender and age pairing is summarized in Table 12.2.7

Table 12.2.7 Contingency table for variables age and gender in clinical trial study data in Table 12.2.4.

		X_{Age}	
		1	0
X_{Gen}	1	$a = 2$	$b = 0$
	0	$c = 0$	$d = 2$

From 12.2.7, the Pearson product moment correlation between variables age and gender is

$$r = \frac{ad - cb}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} = \frac{4}{\sqrt{2 \times 2 \times 2 \times 2}} = 1$$

The result $r = 1$ indicates a perfect similarity. One can interpret this as a higher dependency between two variables and may combine them to form a cluster due to evident similarity. However, it is clear that having only four observations is not conclusive, and the example we consider here is mainly to explain and illustrate the various concepts.

12.3 HIERARCHICAL CLUSTERING METHODS

Even with a moderate amount of data, finding possible clusters without specifying the number of clusters is challenging. In such situations, we have to look for every single possibility of grouping objects into clusters, while determining the efficient number of clusters. Hierarchical clustering procedures are usually helpful in such situations, as they use sequential iterative approaches to build clusters. Hierarchical clustering algorithms are categorized into two main branches, *agglomerative* algorithms and *divisive* algorithms.

The agglomerative hierarchical approaches begin by assuming all the objects are distinct clusters at the initial stage. Then, subsequently, objects are merged into the most appropriate initial clusters, based on a predefined similarity measure. The process continues until all the objects group together to form a one final cluster. Similarity of objects within the clusters at the bottom of the hierarchy would be higher as these are the initial clusters, but similarity naturally decreases as we move up in the hierarchy. The divisive hierarchical approach works completely the opposite way, as it begins with one single cluster of all the objects, and subsequently splits it into two most dissimilar clusters. The process continues until each object falls into its own cluster such that no further divisions are possible. However, the agglomerative approach is known to be more applicable than the divisive approach and therefore, we will discuss that aforementioned approach in detail.

The *Linkage* methods are particularly useful in agglomerative approaches. The major types of linkage methods used in hierarchical clustering procedures include *single linkage*, *complete linkage*, and *average linkage* methods. Single linkage uses the minimum distance between objects in distinct clusters, while complete linkage uses the maximum distance. The average linkage uses the average distance between two distinct clusters. In all the three methods, the minimum intra-class distance criteria should be used to merge clusters, as explained in the following algorithm.

Suppose we have N objects to cluster at the initial step, and that we select a certain linkage method with distance measure $D_{i,j}$ between ith and jth objects.

Step 1 Calculate all pairwise distances $D_{i,j}$ and tabulate.
Step 2 Search the table in Step 1 for $\min(D_{i,j})$ for $i \neq j$ and merge objects (or clusters) that have minimum distance.
Step 3 Delete the corresponding row and the column for the objects (or clusters) previously merged (say A and B) and include the newly formed cluster AB and calculate the distance $D_{AB,k}$ for all k's.
Step 4 Repeat the process until no further merges are possible.

The inal clustering steps, including at what stage they merge may be displayed in a tree-like graphical display called a *Dendrogram.*

12.3.1 Single Linkage

Let us consider two clusters A and B. Then the distance between these two clusters may be defined as

$$D_{A,B} = \min\{D_{x,y} | x \in A, y \in B\} \tag{12.3.1}$$

As stated in Equation (12.3.1), $D_{A,B}$ measures the *minimum distance* or *nearest neighbor distance* between two clusters A and B, as shown below in Figure 12.3.1. Then in the single linkage procedure, the two clusters with the minimum nearest neighbor distance are merged.

Example 12.3.1 (Crime data in Example 12.2.1) *Let us consider the crime data shown in Table 12.2.1 of Example 12.2.1. The Table 12.3.1 shows the calculated Euclidean crime*

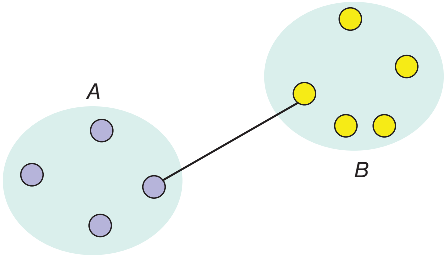


Figure 12.3.1 Single linkage distance between two clusters A and B.

distances between A, B, C, D, and E. Since the distances are symmetric around the diagonal, only the lower triangular distances are displayed.

Table 12.3.1 Euclidean distances between six US states.

	A	B	C	D	E
B	38.5				
C	61.0	80.3			
D	37.8	44.6	36.3		
E	41.7	57.6	23.2	15.0	
F	51.1	59.9	24.5	15.9	**13.9**

The distance between states E and F is the smallest among all the pairwise distances reported in Table 12.3.1. Therefore, we merge those two states to form the initial cluster EF. Then, we calculate the distances between newly formed cluster EF and the rest of the states as follows:

$$D_{EF,A} = \min\,(D_{E,A},\ D_{F,A}) = \min\,(41.7, \quad 51.1) = 41.7$$
$$D_{EF,B} = \min\,(D_{E,B},\ D_{F,B}) = \min\,(57.6, \quad 59.9) = 57.6$$
$$D_{EF,C} = \min\,(D_{E,C},\ D_{F,C}) = \min\,(23.2, \quad 24.5) = 23.2$$
$$D_{EF,D} = \min\,(D_{E,D},\ D_{F,D}) = \min\,(15.0, \quad 15.9) = 15.0$$

The resulting updated distances are summarized as follows in Table 12.3.2.

Table 12.3.2 Updated Euclidean distances after merging E and F.

	A	B	C	D
B	38.5			
C	61.0	80.3		
D	37.8	44.6	36.3	
EF	41.7	57.6	23.2	**15.0**

The distance Table 12.3.2 provides the updated distances and shows that the smallest distance is now between the cluster EF and state D so that we merge them together to form the second cluster *DEF*. The Table 12.3.3 gives updated Euclidean distances based on the following distance calculations:

$$D_{DEF,A} = \min\,(D_{EF,A}, D_{D,A}) = \min\,(41.7,\ 37.8) = 37.8$$

$$D_{DEF,B} = \min\,(D_{EF,B}, D_{D,B}) = \min\,(57.6,\ 44.6) = 44.6$$

$$D_{DEF,C} = \min\,(D_{EF,C}, D_{D,C}) = \min\,(23.2,\ 36.3) = 23.2$$

Table 12.3.3 Updated Euclidean distances after merging *EF* and *D*.

	A	*B*	*C*
B	38.5		
C	61.0	80.3	
DEF	37.8	44.6	**23.2**

It is now clear from the Table 12.3.3 that the state *C* qualifies to merge with cluster *DEF* as their intra-cluster distance (23.2) is the lowest among all. Therefore, in the next stage, we merge state *C* with cluster *DEF* to form the third cluster *CDEF*. Then, we update the distance table based on the following distance calculations as summarized in Table 12.3.4.

$$D_{CDEF,A} = \min\,(D_{DEF,A}, D_{C,A}) = \min\,(37.8,\ 61.0) = 37.8$$

$$D_{CDEF,B} = \min\,(D_{DEF,B}, D_{C,B}) = \min\,(44.6,\ 80.3) = 44.6$$

Table 12.3.4 Updated Euclidean distances after merging *DEF* and *C*.

	A	*B*
B	38.5	
CDEF	**37.8**	44.6

Based on the reported distances in Table 12.3.4, we merge state *A* with the cluster *CDEF* to form the next cluster *ACDEF*. At the final stage, this cluster should merge with the object *B* at the associated distance $D_{ACDEF,B} = 38.5$ to provide the final cluster $BACDEF$.

Figure 12.3.2 shows the single linkage-based *Dendrogram* for the US arrests data for the six states we considered in Example 12.3.1. For instance, if we desire to finalize clusters at distance 20 (see dashed line in Figure 12.3.2), then the algorithm produces four distinct clusters (EFD), (C), (A), and (B). However, if one decides to finalize clusters at distance 30 (see dotted line in Figure 12.3.2), then the algorithm produces three distinct clusters $(EFDC)$, (A), and (B). In this manner, the subjective knowledge may help one to decide at what distance level clusters should be finalized.

Example 12.3.2 (Crime data in Example 12.2.1) *Let us consider the crime data shown in Table 12.2.1 of Example 12.2.1. Obtain the Euclidean distance matrix for all six states,*

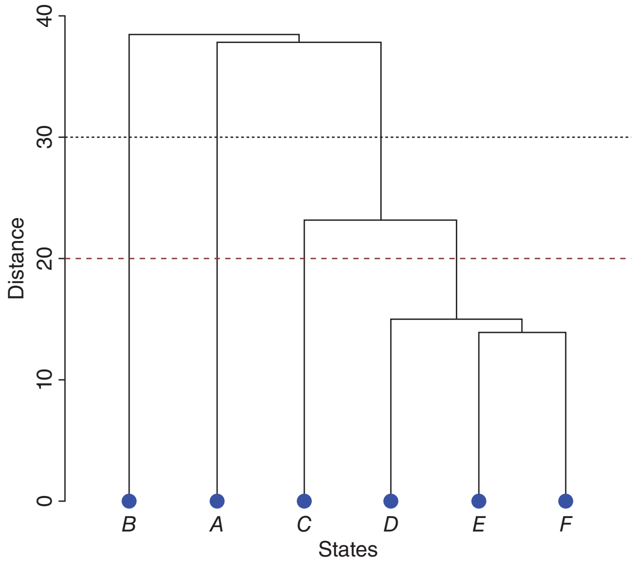


Figure 12.3.2 Dendrogram for US arrests data for six states based on the single linkage. The dashed and dotted lines are manually inserted.

use the single linkage method to cluster A, B, C, D, and E and obtain the dendrogram using both R and MINITAB.

Solution

USING R

The R function ‘hclust()’ can be used to conduct the required cluster analysis as shown in the following R code. The linkage method needs to be specified in this function to obtain the required cluster type.

```
#Data preparation
Murder = c(15.4,13,9,11.3,8.1,11.4)
Assault = c(335,337,276,300,294,285)
Rape = c(31.9,16.1,40.6,27.8,31,32.1)
Urban = c(80,45,91,67,80,70)
crime = cbind(Murder, Assault, Rape, Urban)
row.names(crime) = c(“A”,“B”,“C”,“D”,“E”,“F”)

#Calculate the distance, the following will produce a distance matrix similar to
Table 12.3.1
distance = dist(crime, method = “euclidean”)
distance

#Hierarchical clusters with single linkage option (method option can change as needed)
```

```
single.hc = hclust(distance, method = “single”)

#Obtain dendrogram object for better plots
den.obj = as.dendrogram(single.hc)

#Define node parameters
nodes = list(lab.cex = 1, pch = c(NA, 19), cex = 2, col = “blue”)
plot(den.obj, xlab = “States”, ylab = “Distance”, nodePar = nodes)
abline(h=20, lty=2, col=“red”)
abline(h=30, lty=3, col=“blue”)
```

The R function ‘dist()’ provides the distance table that is identical to Table 12.3.1. The R function ‘hclust()’ provides the required hierarchical clusters, which may be converted to a dendrogram object using the R function ‘as.dendrogram()’. After adding extra graphing arguments via ‘nodePar’ as shown in the above R code, we obtain the dendrogram in Figure 12.3.2.

MINITAB

1. Enter the data in column C1-C5 of the Worksheet and name them State, Murder, Assault, Rape,and Urban, respectively.
2. From the Menu bar, select **Stat** > **Multivariate** > **Cluster Observations** . . .
3. In the new dialog box that appears, enter C2-C5 in the box under **Variables or distance matrix:**, select Single from the menu next to **Linkage Method:**, then select Euclidean from the menu next to **Distance measure:**. Enter 1 in the box next to **Number of clusters:** and then click the check box next to **Show dendrogram**.
4. Click on the **Customize** . . . option and enter C1 in the box next to **Case labels:** select Distance for **Label Y axis with** in the new dialog box that appears. Click **OK** twice.

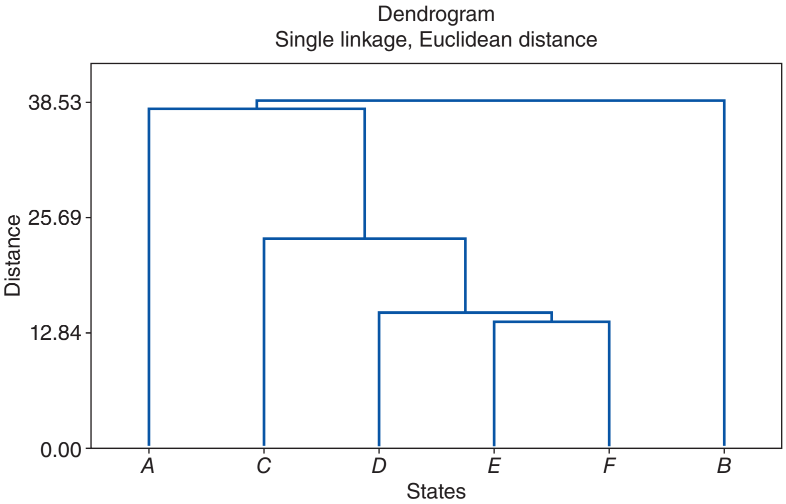


Figure 12.3.3 MINITAB dendrogram plot for US crime data for six states based on the single linkage method.

The MINITAB dendrogram in Figure 12.3.3 provides exactly the same information as obtained by using R (see Figure 12.3.2). However, there is no straightforward way to calculate the initial distance table in MINITAB.

12.3.2 Complete Linkage

Let us consider two clusters A and B. Then, the distance between these two clusters may be defined as

$$D_{A,B} = \max\ \{D_{x,y} | x \in A,\, y \in B\} \tag{12.3.2}$$

As defined in Equation (12.3.2), $D_{A,B}$ measures the *maximum* or the *farthest neighbor distance* between two clusters A and B. Figure 12.3.4 shows the complete linkage distance between two clusters A and B. Then in the complete linkage procedure, we merge the two clusters with the minimum farthest neighbor distance comparing all the farthest neighbor distances.

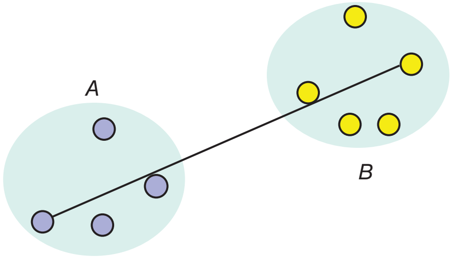


Figure 12.3.4 Complete linkage distance between two clusters A and B.

Example 12.3.3 (Crime data in Example 12.2.1) *We consider the crime data in Example 12.2.1 and the previously calculated Euclidean crime distances between A, B, C, D, E and F shown in Table 12.3.5. Use this data to cluster states using the complete linkage method.*

Table 12.3.5 Euclidean distances between six US states.

	A	B	C	D	E
B	38.5				
C	61.0	80.3			
D	37.8	44.6	36.3		
E	41.7	57.6	23.2	15.0	
F	51.1	59.9	24.5	15.9	**13.9**

States E and F provide the smallest distance, so that these two are merged to form the initial cluster EF as we did in the single linkage method. Then the distances between the newly formed cluster EF and the rest of the states are as follows.

$$D_{EF,A} = \max\ (D_{E,A},\ D_{F,A}) = \max\ (41.7,\ 51.1) = 51.1$$
$$D_{EF,B} = \max\ (D_{E,B},\ D_{F,B}) = \max\ (57.6,\ 59.9) = 59.9$$
$$D_{EF,C} = \max\ (D_{E,C},\ D_{F,C}) = \max\ (23.2,\ 24.5) = 24.5$$
$$D_{EF,D} = \max\ (D_{E,D},\ D_{F,D}) = \max\ (15.0,\ 15.9) = 15.9$$

The following Table 12.3.6 shows the updated distances after merging states *E* and *F*, and using the above four results.

Table 12.3.6 Updated Euclidean distances after merging states *E* and *F*.

	A	*B*	*C*	*D*
B	38.5			
C	61.0	80.3		
D	37.8	44.6	36.3	
EF	51.1	59.9	24.5	**15.9**

It is clear from the above Table 12.3.6 that the cluster *EF* and state *D* provides the minimum distance. Therefore, we merge them to form the second cluster *DEF* and accordingly, we update the distances after completing the following distance calculations as shown in Table 12.3.7.

$$D_{DEF,A} = \max\,(D_{EF,A},\ D_{D,A}) = \max\,(51.1,\ 37.8) = 51.1$$
$$D_{DEF,B} = \max\,(D_{EF,B},\ D_{D,B}) = \max\,(59.9,\ 44.6) = 59.9$$
$$D_{DEF,C} = \max\,(D_{EF,C},\ D_{D,C}) = \max\,(24.5,\ 36.3) = 36.3$$

Table 12.3.7 Updated Euclidean distances after merging cluster *EF* and state *D*.

	A	*B*	*C*
B	38.5		
C	61.0	80.3	
DEF	51.1	59.9	**36.3**

Then we merge the state *C* with *DEF* in the next stage as they provide the minimum intracluster distance to form the third cluster *CDEF*. Then, we update the distances after completing the following distance calculations as shown in Table 12.3.8.

$$D_{CDEF,A} = \max\,(D_{DEF,A},\ D_{C,A}) = \max\,(51.1,\ 61.0) = 61.0$$
$$D_{CDEF,B} = \max\,(D_{DEF,B},\ D_{C,B}) = \max\,(59.9,\ 80.3) = 80.3$$

Table 12.3.8 Updated Euclidean distances after merging cluster *DEF* and state *C*.

	A	*B*
B	**38.5**	
CDEF	61.0	80.3

In the next stage, object A merges with object B as they provide the minimum intracluster distance to form the next cluster AB. At the final stage, this cluster AB should merge with the cluster $CDEF$ at the following distance

$$D_{CDEF,AB} = \max\left(D_{CDEF,A},\ D_{CDEF,B}\right) = \max\left(61.0,\ 80.3\right) = 80.3$$

Figure 12.3.5 shows the complete linkage-based dendrogram for the US arrests data for the six states considered.

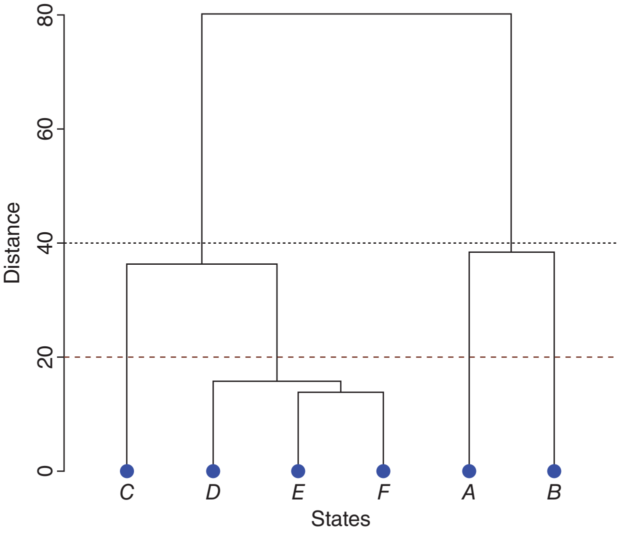


Figure 12.3.5 Dendrogram for the US arrests data for six states based on the complete linkage. The dashed and dotted lines are manually inserted.

For instance, if we desire to finalize clusters at distance 20 (see dashed line in Figure 12.3.5), then the algorithm produces four distinct clusters (EFD), (C), (A), and (B). However, if one decides to finalize clusters at distance 40 (see dotted line in Figure 12.3.5), then the algorithm produces two distinct clusters $(EFDC)$ and (AB). So that one may have to experiment with few different distance values before finalizing the final clusters. Indeed, as explained earlier, the subjective knowledge may help one to decide at what distance level clusters should be finalized.

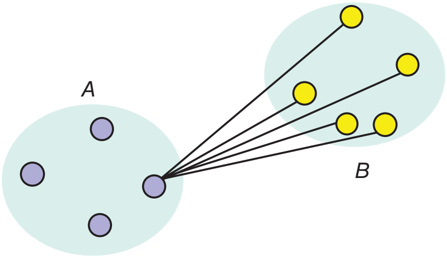


Figure 12.3.6 Average linkage distance between two clusters A and B.

12.3.3 Average Linkage

The average linkage procedure calculates the intracluster distance by averaging all individual intraobject distances between clusters. Then the two clusters with the minimum average distance are merged. For two different clusters as shown in Figure 12.3.6, we calculate the intracluster distance as follows

$$D_{A,B} = \frac{1}{n_A n_B} \sum_{x \in A} \sum_{y \in B} D_{x,y} \tag{12.3.3}$$

where n_A, n_B are the number of objects in cluster A and B, respectively.

Example 12.3.4 (Face clusters) *In this example, we consider a two-dimensional data set that mimics a face-like structure shown in Figure 12.3.7a. There are six apparent clusters in this data set. Both R and MINITAB are used to obtain single, complete, and average linkage method based clusters.*

Note*: We can either generate the data using the 'library(fpc)' and R-code: face = rFace(400, p = 2, nrep.top = 1, smile.coef = .9, dMoNo = 2, dNoEy = 0) or download Example 12.3.4 data from the website: www.wiley.com/college/gupta/statistics2e .*

Solution: We will first use R software to obtain the required clusters.

USING R

The R function 'hclust()' in R can be used to conduct the required cluster analysis as shown in the following R-code. The linkage method needs to be specified in this function to obtain the required cluster type. Note that we can directly generate *Face data set* as required using the 'rFace()' function as shown below.

```
library(stats)
library(fpc)

#Generate the ‘Face’ data set as follows or alternatively place your Example 12.3.4
data in a local folder and import.
set.seed(123)
face = rFace(400, p = 2, nrep.top = 1, smile.coef = .9, dMoNo = 2, dNoEy = 0)
memb1 = as.integer(attr(face,“grouping”))
data = cbind(face[,1], face[,2], memb1)
op = par(mfcol = c(2, 2))

#Plot original Face data
par(las =1)
plot(data, col = as.integer(memb1), pch = as.integer(memb1), xlab=“(a)”,
ylab=“”, main = “True Groups”)
#Plot compete linkage results
hc = hclust(dist(data), method = “complete”)
```

```
memb3 = cutree(hc, k = 6)
plot(data, col = memb3, pch = as.integer(memb3), xlab = "(c)",
ylab = "", main = "Complete Linkage")

#Plot single linkage results
hc = hclust(dist(data), method = "single")
memb2 = cutree(hc, k = 6)
plot(data, col = memb2, pch = as.integer(memb2), xlab = "(b)",
ylab = "", main = "Single Linkage")

#Plot average linkage results
hc = hclust(dist(data), method = "average")
memb4 = cutree(hc, k = 6)
plot(data, col = memb4, pch = as.integer(memb4), xlab = "(d)",
ylab = "", main = "Average Linkage")
```

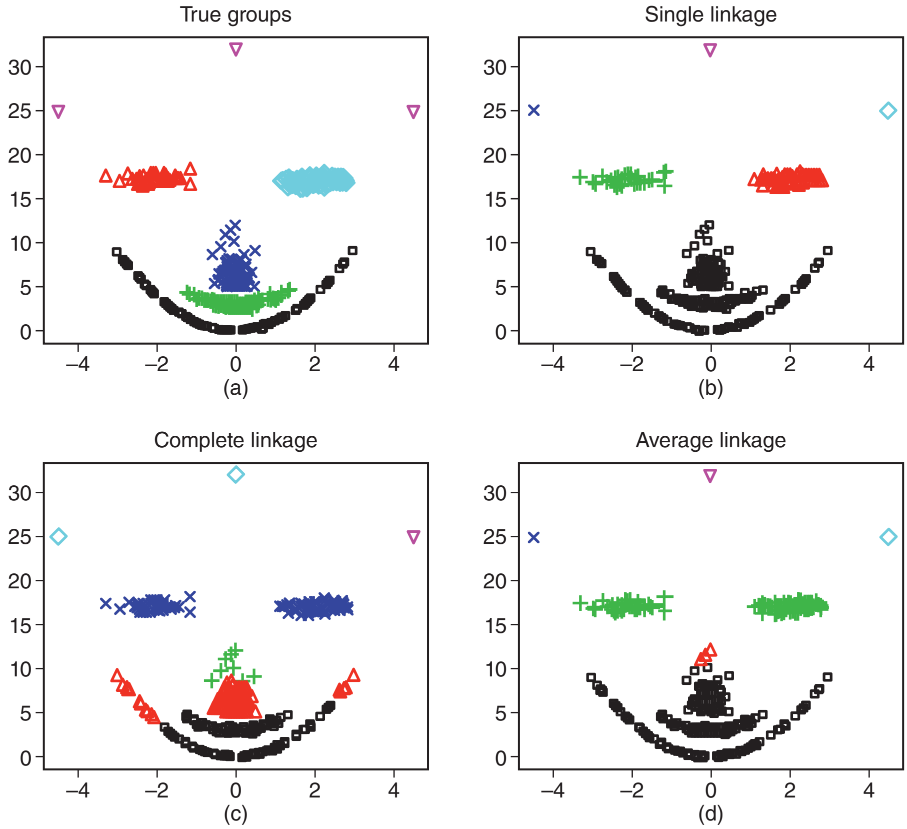


Figure 12.3.7 A generated face structure with six different clusters and their estimates by using single, complete, and average linkage methods. Different symbols (colors) represent different clusters.

This specific face structure (see Figure 12.3.7a) includes chains, elliptical and triangular-shaped clusters as well as outliers. It seems that this is a relatively difficult data set to cluster. Nevertheless, we expect at least to differentiate major objects of the face structure, such as the two eyes, nose, and chin into different clusters. Figure 12.3.7b shows that the single linkage method accurately clusters eyes, but fails to differentiate the nose, mouth, and chin, and forms a large single clump. One possible reason is the presence of the chain-like chin. However, the complete linkage method accurately identifies the mouth and nose but combines the lower part of the chin with the mouth, and the upper part of the chin with the nose (see Figure 12.3.7c). Hence, it makes sense to split the long chin into a few separate clusters, since the data on the two extreme ends tend to separate from the middle part. Also, the complete linkage method tends to produce clusters with somewhat similar dimensions. We cannot see much improvement on the average linkage results, since this method fails to distinguish the two eyes accurately.

The single linkage method tends to generate "long chains" and "clumps" like clusters and handles well nonelliptical shapes, but is sensitive to outliers and noise and may not produce meaningful results in the presence of complex structures. The complete linkage method tends to produce more balanced clusters and may not be affected by outliers as much as the single linkage method does. However, it splits the large clusters into small clusters, while nearby small clusters tend to merge with large clusters. The average linkage results (see Figure 12.3.7d) tend to compromise between the single and complete linkage outcomes. In addition, it is less susceptible to noise and outliers.

MINITAB

1. Enter the data in column C1 and C2 of the Worksheet and name them X and Y, respectively.
2. From the Menu bar, select **Stat** > **Multivariate** > **Cluster Observations** ...
3. In the new dialog box appears, enter C1 and C2 in the box under **Variables or distance matrix:**, select any linkage method from the menu next to **Linkage Method:** and select Euclidean from the menu next to **Distance measure:**. Enter 6 in the box next to **Number of clusters:**.
4. Click **Storage** ... option and enter C3 in the box next to **Cluster membership column:** a new dialog box that appears.
5. Click **OK** twice. A complete output will appear in the session window. However, to plot clusters, we proceed as follows:
6. From the Menu bar select **Graph** > **Scatterplot** .... This prompts a dialog box to appear on the screen. In this dialog box, select Scatterplots **With Groups** and click **OK**. In the new dialog box that now appears, under the X and Y variables, enter C1 and C2, respectively. Use the desired options and click **OK**. A scatter plot with clustered groups will appear.

12.3.4 Ward's Hierarchical Clustering

Ward's method is another hierarchical clustering procedure that begins with considering every object as a distinct cluster. Objects merge with minimum *merging cost* defines in (12.3.4). Let us consider merging two objects A and B to form a new cluster AB that results in a cluster center $\overline{AB}$, where $\overline{AB}$ represents the central location of objects A and B. Then, the resulting merging cost due to clustering is defined as

$$Cost_{A,B} = D^2_{A,\ \overline{AB}} + D^2_{B,\ \overline{AB}} \tag{12.3.4}$$

where $D^2_{A,\,\overline{AB}}$ and $D^2_{B,\,\overline{AB}}$ denote the squared Euclidean distances from the objects A and B to the center $\overline{AB}$ of objects A and B, respectively. It quantifies the within cluster sum of squared deviation.

At the initial step in the Ward's method, we merge object A with a distinct object B as long as it provides the minimum merging cost among all possible such pairwise merging costs. Prior to the initial step, the total cost is zero, and the cost grows as we merge the objects and clusters. However, the Ward's method keeps the growth of merging cost as minimum as possible.

Example 12.3.5 (Face clusters) *Apply Ward's clustering method for the data in Example 12.3.4 and discuss its results.*

Solution: We will illustrate Ward's clustering method using R as shown below.

USING R

As in the previous Example 12.3.4, the 'hclust()' function in R is used to conduct the required cluster analysis.

```
#Continued from the previous R code in Example 12.3.4
op = par(mfcol = c(1, 2))
par(las =1)

#Original cluster data
plot(data, col = as.integer(memb1), pch = as.integer(memb1),
xlab="(a)", ylab="", main = "True Groups")

#Ward's clusters
hc = hclust(dist(data), method = "ward.D2")
memb5 = cutree(hc, k = 6)
plot(data, col = memb5, pch = as.integer(memb5), xlab = "(b)",
ylab = "", main = "Ward's Method")
```

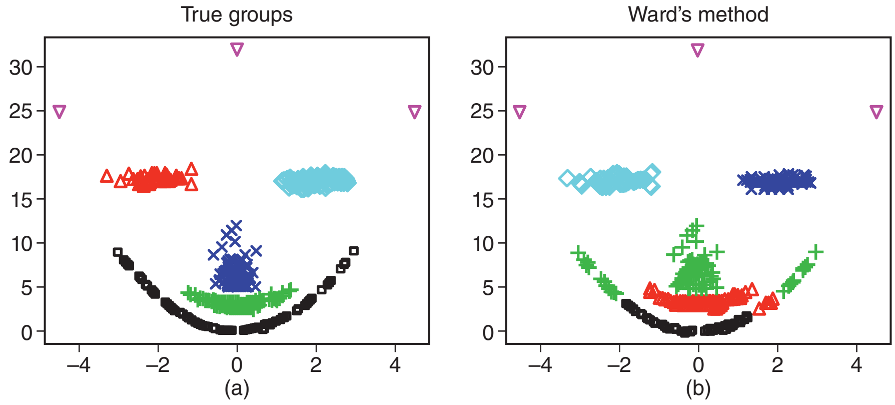


Figure 12.3.8 (a) A generated face structure with six different clusters (b) Ward's method that establishes estimated clusters. Different symbols (colors) represent different clusters.

Now Figure 12.3.8 shows that there is a substantial improvement in the Ward's clustering process compared to the earlier linkage results in Figure 12.3.7. The Ward's method clearly identifies both the eyes, nose, and mouth, but parts of the chin are misclassified and clustered with the mouth and nose. Also, three outliers are grouped into one cluster accurately. However, it is important to note that though the Ward's method tend to produce better results in many situations, but some linkage methods may outperform the Ward's method in some other situations. Therefore, we cannot make any general statement about the performances of these methods.

12.4 NONHIERARCHICAL CLUSTERING METHODS

12.4.1 *K*-Means Method

The *K*-means method is one of the popular nonhierarchical clustering procedures. It requires a user to define the value of K the number of desired clusters. The procedure begins with assigning K initial centroids. In general, the centroid can either be mean or median distance between a set of objects. In *K*-means method, we consider the mean as the centroid. In this method, once the centroids are identified, each object is assigned to the nearest centroid. As a result, the locations of the centroids of the subsequent clusters should be updated after each assignment. Major steps of the procedure are summarized as follows:

Step 1 Assign K initial centroids. This can be done by either splitting objects into K arbitrary clusters and calculating their centroids or randomly assigning K centroids.

Step 2 Assign objects to their nearest centroids one at a time. In each assignment, the centroid should be updated, and the updated centroids should be used in the proceeding step.

Step 3 Repeat object allocation until the centroids are stable where no further assignment is possible.

In order to assign the objects to clusters, a predefined measure of distance should be used. Then the nearest distance is defined with respect to that distance measure. Euclidean distance is a popular choice in the *K*-means procedure, but the appropriate similarity measure should be used, as discussed in Section 12.2. In the cases where a more robust centroid measure is needed, we may use the median. However, adapting median centroid along with the city-block distance that utilizes the absolute-error criterion in the above algorithm is called the *K*-medoids method (see Review problems 12.12 and 12.13 for details).

Example 12.4.1 *Consider a two-dimensional data set that consists of seven objects labeled $A - G$ as shown in Table 12.4.1. We suppose that these seven objects belong to two distinct clusters ($K = 2$). Use the K-means method to cluster these seven objects.*

Table 12.4.1 A two-dimensional data set.

Labels	A	B	C	D	E	F	G
X	1	2	3	0	2	4	3
Y	1	3	4	1	5	4	2

Solution: Since there are no information available about possible centroids, we split the data into two arbitrary clusters such that objects ADG and $BCEF$ belong to two initial clusters. Then, we can calculate the centroids (means) of the clusters as follows:

$$(\overline{X}_{ADG}, \overline{Y}_{ADG}) = \left(\frac{1+0+3}{3}, \frac{1+1+2}{3}\right) = (1.33, 1.33)$$

and

$$(\overline{X}_{BCEF}, \overline{Y}_{BCEF}) = \left(\frac{2+3+2+4}{4}, \frac{3+4+5+4}{4}\right) = (2.75, 4.00)$$

These resulting centroids are labeled by symbol (*) in Figure 12.4.1a. Then, we calculate the Euclidean distance from object A to its current and the opposite cluster centroids.

$$D_{A,\,ADG} = \sqrt{(1-1.33)^2 + (1-1.33)^2} = 0.47$$

$$D_{A,\,BCEF} = \sqrt{(1-2.75)^2 + (1-4)^2} = 3.47$$

With respect to the above centroids, object A should belong to the current cluster, as the distance $D_{A,\,ADG}$ is much smaller than $D_{A,\,BCEF}$. However, we still calculate the distances under the assumption that we move object A to the opposite cluster $BCEF$. Thus, we calculate the new cluster centroids of DG and $ABCEF$ due to this hypothetical assignment.

Thus, we have new clusters DG and $ABCEF$, and their centroids are

$$(\overline{X}_{DG}, \overline{Y}_{DG}) = \left(\frac{0+3}{2}, \frac{1+2}{2}\right) = (1.5, 1.5)$$

and

$$(\overline{X}_{ABCEF}, \overline{Y}_{ABCEF}) = \left(\frac{1+2+3+2+4}{5}, \frac{1+3+4+5+4}{5}\right) = (2.4, 3.4)$$

These centroids (*) are shown in Figure 12.4.1b. Resulting distances from object A to new centroids (1.5, 1.5) and (2.4, 3.4) are

$$D_{A,\,DG} = \sqrt{(1-1.5)^2 + (1-1.5)^2} = 0.71$$

$$D_{A,\,ABCEF} = \sqrt{(1-2.4)^2 + (1-3.4)^2} = 2.78$$

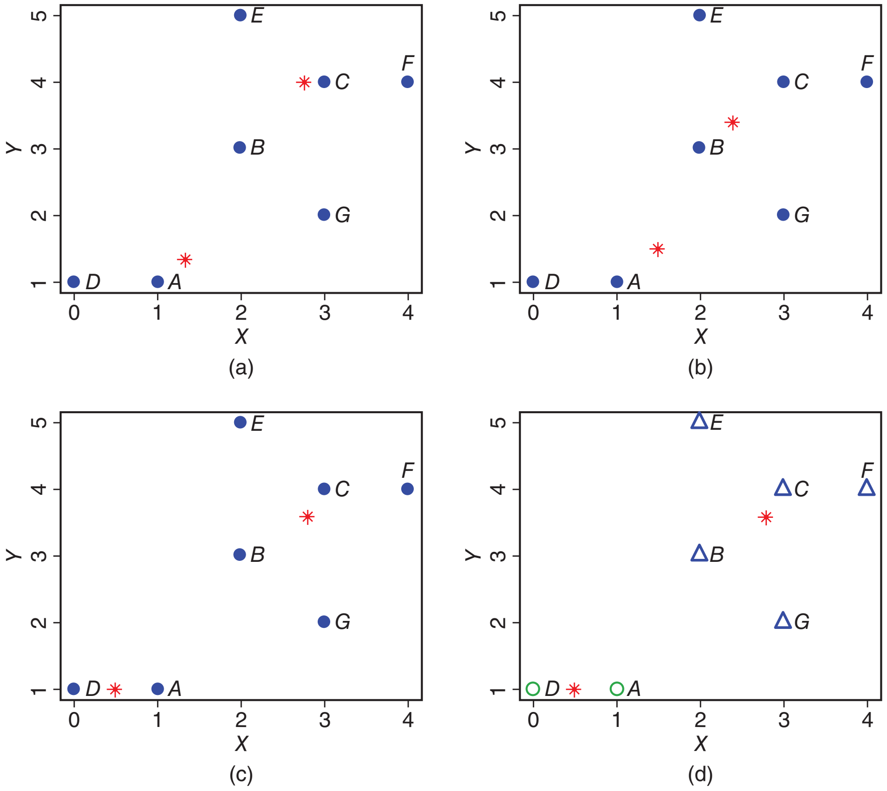


Figure 12.4.1 R output plot of seven two-dimensional objects. The cluster centroids of each cluster is represented by a "*." Final two clusters are shown in (d) using symbols '∘' and '△'.

It is clear from this calculation that the object A should not be moved from its initial cluster ADG to $BCEF$, as it is closer to its initial cluster center with distance $D_{A,ADG} = 0.47 (< D_{A,DG} = 0.71 < D_{A,ABCEF} = 2.78 < D_{A,BCEF} = 3.47)$.

Now let us consider the object G and its distances with respect to the initial cluster centroids $(1.33, 1.33)$ and $(2.75, 4.00)$, since we fail to move object A.

$$D_{G,ADG} = \sqrt{(3-1.33)^2 + (2-1.33)^2} = 1.80$$

$$D_{G,BCEF} = \sqrt{(3-2.75)^2 + (2-4)^2} = 2.02$$

To recalculate the distances under the assumption that we move object G to opposite cluster $BCEF$, let us first update new centroids.

$$(\overline{X}_{AD}, \overline{Y}_{AD}) = \left(\frac{1+0}{2}, \frac{1+1}{2}\right) = (0.50, 1.00)$$

and

$$(\overline{X}_{BCEFG}, \overline{Y}_{BCEFG}) = \left(\frac{2+3+2+4+3}{5}, \frac{3+4+5+4+2}{5}\right) = (2.80, 3.60)$$

These new centroids (*) are plotted in Figure 12.4.1c. The distances with respect to the new centroids are

$$D_{G,AD} = \sqrt{(3-0.50)^2 + (2-1.00)^2} = 2.69$$

$$D_{G,BCEFG} = \sqrt{(3-2.80)^2 + (2-3.60)^2} = 1.61$$

Since the distance from object G to the centroid $(\overline{X}_{BCEFG}, \overline{Y}_{BCEFG}) = (2.80,\ 3.60)$ is closer than rest of the distances, we should retain object G in its new cluster $BCEFG$.

We continue to reassign the objects to opposite clusters until no further allocation is possible. In other words, the cluster centroids must be stable at the final stage of the K-means algorithm. However, due to growing number of possible object reallocations, we use one of the software discussed in this text to perform the algorithm. In this example, fortunately, cluster centroids are stable after the above reassignment. So that the K-means method produces two distinct clusters AD and $BCEFG$ that are shown on the Figure 12.4.1d and identified via symbols "∘" and "Δ," respectively. The basic summary statistics of both the final clusters are tabulated in Table 12.4.2 and further summary statistics can be calculated using both R and MINITAB software as shown below.

Table 12.4.2 Summary statistics of K-means clusters in Example 12.4.1.

Cluster	Members	Centroid	Within cluster SS
Cluster 1	A, D	(0.50, 1.00)	$(1-.5)^2 + (1-1)^2 + (0-.5)^2 + (1-1)^2 = 0.5$
Cluster 2	$B, C, E,$ F, G	(2.80, 3.60)	$(2-2.8)^2 + (3-3.6)^2 + (3-2.8)^2 + (4-3.6)^2$ $+ (2-2.8)^2 + (5-3.6)^2 + (4-2.8)^2 + (4-3.6)^2$ $+ (3-2.8)^2 + (2-3.6)^2 = 8.0$

USING R

To conduct the K-means clustering algorithm in R, we can use the 'kmeans()' function as shown below in the R code.

```
#Prepare data
X = c(1,2,3,0,2,4,3)
Y = c(1,3,4,1,5,4,2)
data = cbind(X, Y)
row.names(data) = c(“A”,“B”,“C”,“D”,“E”,“F”,“G”)


#To obtain K-means clusters
K.means = kmeans(data, 2)
K.means


# R output
K-means clustering with 2 clusters of sizes 5, 2
Cluster means:
     X     Y
  ------------
  1  2.8   3.6
  2  0.5   1.0
  ------------

Clustering vector:

 A   B   C   D   E   F   G
 ------------------------
 2   1   1   2   1   1   1

Within cluster sum of squares by cluster:
[1] 8.0 0.5
(between_SS / total_SS = 66.9 %)
```

The above R output provides the similar information that we have obtained in manual calculations summarized in Table 12.4.2. Note that the cluster centroids are simply named as the cluster means in the above R output. The following R code can be used to visualize the resulting *K*-means clusters shown in Figure 12.4.1d.

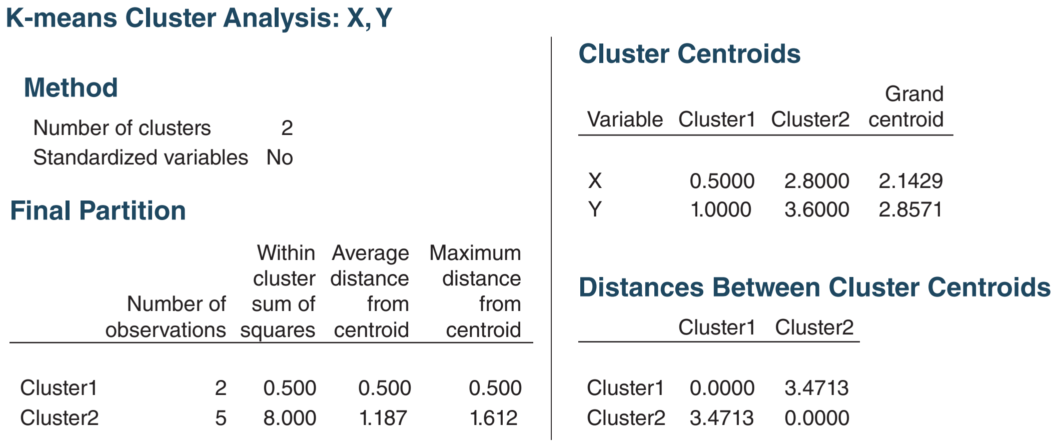
```
#To plot above K-means clusters
plot(Y~X, col = (K.means$cluster+2), pch=c(1,2,2,1,2,2,2), cex=1.5, xlab = “X”, ylab =
“Y”, sub = “(d)”, data=data)
text(Y~X, labels = row.names(data), pos = c(4,4,4,4,4,3,4))

#To plot the center
points(K.means$centers, col=2, pch = 8, lwd=2)
```

The results we obtain from the R ‘kmeans()’ function are similar to that we obtained earlier in manual calculations. Now, we will exhibit MINITAB *K*-means procedure.

MINITAB

1. Enter the data in columns C1, C2, and C3 of the Worksheet and name them as Labels, X, and Y, respectively.
2. From the Menu bar, select **Stat** > **Multivariate** > **Cluster K-Means** ...
3. In the new dialog box appears, enter C2 and C3 in the box under **Variables:**, enter 2 in the box next to **Number of clusters:**.
4. Click **Storage** ... option and enter C4 in the box next to **Cluster membership column:** and C5 C6 in the box below **Distance between observations and cluster centroids:** in the new dialog box that appears.

5. Click **OK** twice. The following results will appear in the session window:



```
K-means Cluster Analysis: X, Y

 Method

  Number of clusters         2
  Standardized variables    No

 Final Partition

                          Within   Average    Maximum
                          cluster  distance   distance
               Number of  sum of       from       from
            observations  squares  centroid   centroid

 Cluster1              2    0.500     0.500      0.500
 Cluster2              5    8.000     1.187      1.612

Cluster Centroids

                                          Grand
 Variable  Cluster1  Cluster2   centroid

 X           0.5000    2.8000     2.1429
 Y           1.0000    3.6000     2.8571

Distances Between Cluster Centroids

           Cluster1  Cluster2

 Cluster1    0.0000    3.4713
 Cluster2    3.4713    0.0000
```

6. To plot the clusters, we proceed as follows:
7. From the Menu bar select **Graph** > **Scatterplot** .... This prompts a dialog box to appear on the screen. In this dialog box, select Scatterplots **With Groups** and click **OK**. In the new dialog box that appears, under the X and Y variables, enter C2 and C3, respectively. Enter C4 in the box below **Categorical variables for grouping (0-3):** and click the **Labels** ... option, and in the new dialog box that appears select **Data Labels**, and select the **Use labels from column:** and enter C1. Click **OK** twice. A scatter plot with clustered groups will appear.

As we expected, the above MINITAB output (see results obtained in Step 5) provides two distinct clusters with the cluster centroids located at $(0.50, 1.00)$ and $(2.80, 3.60)$. Also, it provides some additional summary statistics such as within cluster sum of squares, average distance from centroids, and distance between cluster centroids. It is clear from the above MINITAB output that its results are similar to what we observed in both manual (see Table 12.4.2) and the above R calculations. Also, the MINITAB cluster output plot of 7 two-dimensional objects shown in Figure 12.4.2 is similar to what we observed in R cluster output plot shown in Figure 12.4.1d.

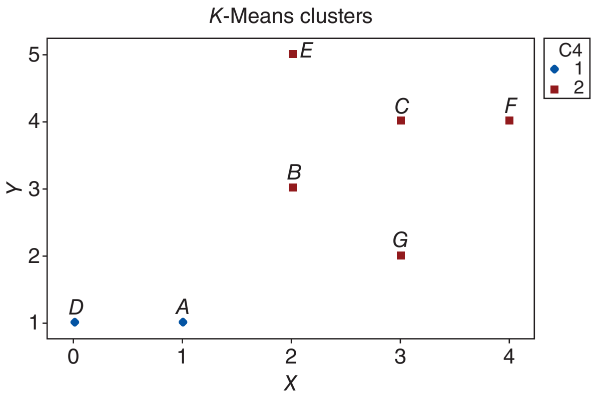


Figure 12.4.2 MINITAB cluster output plot of seven two-dimensional objects.

12.5 DENSITY-BASED CLUSTERING

In this section, we discuss the density-based clustering algorithm DBSCAN which stands for density-based spatial clustering of applications with noise. It uses local densities of the spatial region to form clusters. It tends to separate high-density regions from the low-density regions while connecting similar high-density regions into clusters based on their proximities. It defines a cluster as a maximal (or largest possible) set of density-connected points. This method provides clusters of arbitrary shapes, which is resistant to outliers.

Before we explain the DBSCAN algorithm, it is important to define its parameters and some related quantities.

ϵ-neighborhood: The neighborhood within a radius ϵ of an object of interest.

$N_\epsilon(p)$: A collection of set of objects within ϵ-neighborhood of an object p

$$N_\epsilon(p) = \{q \,|\, d(p,q) \leq \epsilon\}$$

m: Minimum number of objects that are within an ϵ-neighborhood of an object.

Core object: An object is called a *core object* if $|N_\epsilon(p)| \geq m$, where $|N_\epsilon(p)|$ indicates the number of objects in $N_\epsilon(p)$. In other words, the core objects contain at least the predefined minimum number of objects (m) in its ϵ-neighborhood.

Border object: An object is called a *border object* if $|N_\epsilon(p)| < m$ (i.e. it contains less than m objects) but it is within the ϵ-neighborhood of a core object.

Noise object: An object is called a *noise object* if it is neither core nor border object.

Directly density-reachable: If an object q belongs to the ϵ-neighborhood of a core object p (i.e. $q \in N_\epsilon(p)$), then the object q is *directly density-reachable* from object p.

Density-reachable: An object p is *density-reachable* from object q with respect to ϵ and m if there is a chain of objects $p_1, \ldots, p_n$,where $p_1 = q$, and $p_n = p$ such that p_{i+1} is directly density-reachable from p_i.

Density-connected: An object p is *density-connected* to object q with respect to ϵ and m, if there is an object r such that both p and q are density-reachable from r with respect to ϵ and m.

Density-reachability is a result of a sequence of directly density-reachable objects. Neither density-reachability nor direct density-reachability is a symmetric property since no point can be reachable from a noncore point though a noncore point may be reachable from a core point. However, the density-reachability among core objects shows symmetry. By the definition, the density-connectivity is symmetric since if object p is density-connected to object q, then object q is also density-connected to object p.

Example 12.5.1 (Prototype DBSCAN example) *This example includes a set of two dimensional objects shown in Figure 12.5.1. We use these set of objects to exhibit basic*

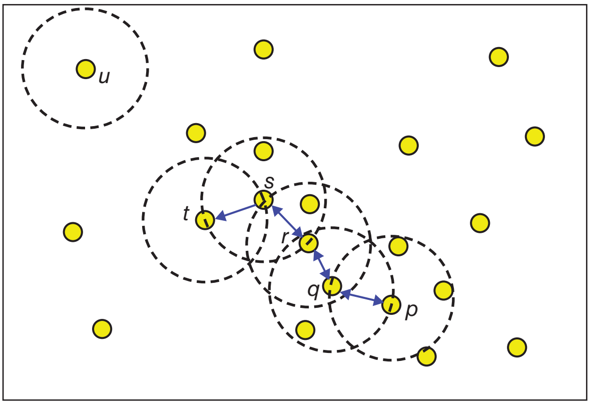


Figure 12.5.1 DBSCAN process for two-dimensional objects.

terminologies in the DBSCAN algorithm including density-reachability and density connectivity. Let $\epsilon = 1$ unit and $m = 4$.

Figure 12.5.1 shows a set of two dimensional objects along with few circular neighbourhoods of radius 1 ($\epsilon = 1$) for few selected objects. As shown in Figure 12.5.1, the object u is a low-density point as $N_\epsilon(u) = 1$, and it is a noise object as it does not belong to a neighborhood of any core object. Objects p, q, r, and s are core objects as each contains at least four neighborhood objects. Object t belongs to the neighborhood of the core object s, but its neighborhood has only two objects and therefore, it is a border object.

Object t is directly density-reachable from object s but object s is not directly density-reachable from object t because the object t is not a core object. Object s is indirectly density-reachable from object p and object p is indirectly density-reachable from object s. Objects p, r, and t are all density connected. Therefore, the objects p, q, r, s, and t belong to the same cluster.

DBSCAN algorithm can be summarized as follows:

1. Select an arbitrary point p.
2. Search p's ϵ-neighborhood and if $N_\epsilon(p) \geq m$, then a new cluster is formed with p as a core object, otherwise p is a border or a noise object.
3. If p is a core object, then collect all density-reachable objects from p to extend the current cluster.
4. If p is a border object, then it has no density-reachable objects, and therefore a previously nonvisited new object is selected.
5. The process terminates when no further objects in the database can be added to any cluster.

A cluster that is produced by a DBSCAN algorithm is the largest possible set of density-connected objects with respect to density-reachability. Therefore, objects in two separate clusters are not density reachable, but if they are then those two clusters should merge into a single cluster, though they have different local densities. Any object that does not belong to a cluster is considered a noise object.

DBSCAN method has several advantages. Its algorithm does not require prespecifying the number of clusters. In general, it produces arbitrarily shaped clusters, which are self-contained. The algorithm identifies noisy objects and therefore it is resistant to outliers. There are, however, a few disadvantages of this algorithm. Confusion may occur when selecting border objects, since such objects could be density-reachable from two separate clusters, depending on the order of the objects processed. Highly ranging local densities may cause issues due to difficulty of choosing DBSCAN parameters in such situations. As in many of the distance-based algorithms, the quality of DBSCAN results is also get affected by the choice of the distance metric.

Example 12.5.2 *We reconsider the face clustering discussed in Example 12.3.4 and apply the DBSCAN algorithm using R and MINITAB.*

USING R

The function ‘dbscan()’ in R can be used to execute the DBSCAN algorithm as shown in the following R code by setting $\epsilon = 0.5$ and $m = 5$.

```
library(dbscan)
#Import face data as explained in Example 12.3.4 data

#Make the data matrix with required scaling
data = cbind(face[,1], face[,2]/2.5)

#Run ‘dbscan()’ function and plot the results
dbs = dbscan(data, eps = .5, minPts = 5)
plot(data, col = dbs$cluster + 1L, pch = as.integer(dbs$cluster), xlab =" ",
ylab =" ", main = " ")
```

Figure 12.5.2 shows the results of the DBSCAN algorithm with parameters $\epsilon = 0.5$ and $m = 5$. The results are comparatively better as it accurately identifies two eyes, nose, and mouth as four distinct clusters. The user-defined outliers are identified as noisy objects, and a few more bordering observations are classified as outliers near

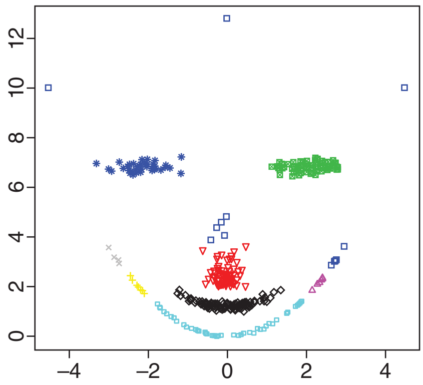


Figure 12.5.2 DBSCAN-based clustering results for face clustering data.

the nose and chin. The long chin breaks into five different groups due to low densities at both the ends of the chin. It is evident that this method fails to cluster long chain-like structures accurately in the presence of low-density bordering and noisy points. However, it is better at distinguishing neighboring large clumps.

MINITAB

DBSCAN option is not available in MINITAB.

12.6 MODEL-BASED CLUSTERING

One disadvantage of linkage-based, Ward's, K-means, and DBSCAN clustering methods is that they are mostly heuristic and do not follow any statistical model. Hence, we cannot explain how the data were generated. The model-based clustering method assumes data were generated from a certain statistical model, and it uses a mixture of probability distributions to recover the original model from which data were possibly generated. The model we estimate from the data defines clusters and assignment of objects to clusters. Refer to Banfield and Raftery (1993) and Melnykov and Maitra (2010) for more details.

Assume we observe n multivariate observations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{n}]'$, where each $\mathbf{x}_i$ has p components. Then the data matrix can be written in the following form.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_i' \\ \vdots \\ \mathbf{x}_\mathrm{n}' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}_{n\times p}$$

We consider each multivariate observation as a separate object with p measurements. We consider the ith object $\mathbf{x}_i$ and assume that the probability of this object belonging to the kth cluster is π_k, with $\pi_k \geq 0$ for all $k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$, where K is the number of distinct clusters. Thus, we can model $\mathbf{x}_i$ via a mixture model with K components. That is, the resulting joint distribution of the $\mathbf{x}_i$'s can be written as a mixture of K distributions by assuming each cluster generates data from its own component distribution f_k, where $k = 1, 2, \ldots, K$.

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i), \quad k = 1, 2, \ldots, K \tag{12.6.1}$$

Observations are more often heterogeneous in nature, and therefore using mixture models seem to be more attractive. In general, the component distribution may or may not have the same form. However, multivariate normal distributions are often used as the common mixture components such that $f_k(\mathbf{x}_i)$ is the density function of $\boldsymbol{N_p}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$. A mixture distribution for normal mixtures can be written as

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-(\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)} \tag{12.6.2}$$

This assumes that observations from the kth cluster are centered at the mean vector $\boldsymbol{\mu_k}$, but that the orientation and size of the cluster depends on the variance-covariance matrix $\boldsymbol{\Sigma}_k$. Now, due to the bell shape nature of the normal distribution and elliptical nature of normal contours, this mixture model tends to produce ellipsoidal clusters.

Component estimation of the above mixture distribution in (12.6.2) is complex, though there are many methods available in the literature. We will discuss the most commonly used maximum likelihood approach. The resulting multivariate normal likelihood function for n objects is

$$L(\boldsymbol{\Theta}\,;\boldsymbol{X}) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-(\mathbf{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)} \tag{12.6.3}$$

where $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_K;\ \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K};\ \boldsymbol{\Sigma_1}, \ldots, \boldsymbol{\Sigma_K})$ are the parameters which we estimate in clustering.

The standard maximum likelihood estimation (MLE) method is used to find the estimator $\hat{\boldsymbol{\Theta}}$ of the parameter vector $\boldsymbol{\Theta}$, where $\hat{\boldsymbol{\Theta}} = \text{Argmax } L(\boldsymbol{\Theta}, \boldsymbol{X})$. We note that the complexity of parameter estimation depends on the nature of the $\boldsymbol{\Sigma}_k$.

The *Expectation-Maximization* (EM) algorithm is commonly applied in many software packages to estimate the MLE of $\boldsymbol{\Theta}$. In the cases where the number of clusters is unknown, the *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC) (see Akaike, 1973; Schwarz, 1978) are used to compare models with different number of distinct clusters, while imposing some penalty for number of the model parameters.

After estimating the model parameters of the mixture distribution, we can estimate the probability of cluster membership of objects. The probability π_{ik} of the ith object $\mathbf{x_i}$ belongs to the kth cluster given by the Bayes rule is

$$\pi_{ik} = \frac{\hat{\pi}_i f_k(\mathbf{x_i}|\hat{\theta}_k)}{\sum_{i=1}^{K} \hat{\pi}_i f_k(\mathbf{x_i}|\hat{\theta}_k)} \tag{12.6.4}$$

We evaluate this membership probability of a given object $\mathbf{x_i}$ for all the clusters and assign it to the highest probable cluster.

Due to the mathematical challenges of executing model-based clustering procedures, it is required to use software for this purpose. Table 12.6.1 shows commonly used covariance structures available in the 'mclust' package in R. We will illustrate the R procedure for estimating the appropriate models for a given set of covariance structures.

Table 12.6.1 Covariance structures for model-based clustering.

Covariance structure	Interpretation
$\boldsymbol{\Sigma}_k = \boldsymbol{\lambda I}$, $\boldsymbol{\lambda}$ is a scalar (EII)	All the clusters are spherical with same volume
$\boldsymbol{\Sigma}_k = \boldsymbol{\lambda_k I}$, $\boldsymbol{\lambda_k}$ are scalars (VII)	All the clusters are spherical with different volumes
$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, for all k (EEE)	All the clusters have same shape, orientation, and volume
$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k$ is unique (VVV)	Each cluster has different shape orientation, and volume
$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Sigma}$ (VEE)	All the clusters have same shape, orientation, and different volume

Example 12.6.1 (Stiffness measurements) *In this example, we consider three stiffness measurements* (X_1, X_2, X_3) *measured from 75 different boards. Observations for the first 13 boards are shown in Table 12.6.2, and the complete data set is available on the website: www .wiley.com/ college/ gupta/ statistics2e. Apply the model-based clustering algorithm, with covariance structures that are shown in Table 12.6.1, and discuss their appropriateness.*

Table 12.6.2 Stiffness measurements for first 13 boards.

Board No.	1	2	3	4	5	6	7	8	9	10	11	12	13
X_1	1084	1059	1017	956	992	907	1152	1009	1215	863	1035	1057	999
X_2	977	1118	973	1053	823	951	1141	892	1029	979	964	870	913
X_3	958	845	994	1027	1174	974	1210	1086	939	1064	936	897	1071

USING R

The function ‘Mclust()’ in R can be used to execute the model-based clustering algorithm that is shown in the following R code. In the ‘Mclust()’ function, we can select the model covariance structure using the ‘modelNames’ option. The ‘Mclust()’ function outputs many important summary statistics. For example, we can easily extract the BIC values for model selection purpose as shown below in the R code.

```
#Install mclust library
install.packages(“mclust”)
library(mclust)

#Import data: make sure to place your data in a local folder and import.
data = read.table(“C:/Users/.../Table 12.6.2.txt”, header=TRUE)[,-1]

#Spherical clusters, equal volume
mod1 = Mclust(data, modelNames =“EII”); plot(mod1, what = “classification”)
summary(mod1)

#Spherical clusters, unequal volume
mod2 = Mclust(data, modelNames =“VII”); plot(mod2, what = “classification”)

#Ellipsoidal, equal volume, shape, and orientation
mod3 = Mclust(data, modelNames =“EEE”); plot(mod3, what = “classification”)

#Ellipsoidal, varying volume, shape, and orientation
mod4 = Mclust(data, modelNames =“VVV”); plot(mod4, what = “classification”)

#Ellipsoidal, equal shape and same orientation but varying volume
mod5 = Mclust(data, modelNames =“VEE”); plot(mod5, what = “classification”)

#Obtain BIC values
model = Mclust(data) ; model$BIC
plot(model$BIC[,1], main = “ ”, xlab = “Number of Clusters”, ylab = “BIC”,
type = “b”, lwd = 2, pch = 2, col = 1, ylim = c(-2760, -2655))
```

```
points(model$BIC[,2], type = “b”, lwd = 2, pch = 3, col = 2)
points(model$BIC[,7], type = “b”, lwd = 2, pch = 4, col = 3)
points(model$BIC[,14], type = “b”, lwd = 2, pch = 5, col = 4)
points(model$BIC[,9], type = “b”, lwd = 2, pch = 6, col = 5)
legend(“topright”, c(“EII”, “VII”, “EEE”, “VVV”, “VEE”), pch= c(2,3,4,5,6),
col=c(1:5), lwd =2)
```

Five different covariance structures listed in Table 12.6.1 are used to fit the model-based clusters for this data set. The estimated BIC values against the number of clusters in each selected covariance structure (or model) are plotted in Figure 12.6.1. The BIC method favors the “VEE” covariance structure that assumes $\boldsymbol{\Sigma}_k = \boldsymbol{\lambda}_k \boldsymbol{\Sigma}$ with only two clusters, as it provides the smallest absolute BIC value. It provides ellipsoidal, equal shape and orientation but different volumed clusters. Two-dimensional projections of those “VEE” clusters are shown in Figure 12.6.2a. The “EEE” covariance structure ($\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$) with only two clusters provides the second highest BIC value. As expected, this method provided clusters with ellipsoidal, equal volume, shape, and orientation (see Figure 12.6.2b). The “VII” covariance structure ($\boldsymbol{\Sigma}_k = \boldsymbol{\lambda}_k \boldsymbol{I}$) favors three clusters.

It is clear from the Figure 12.6.2 that the variable X_2 plays a vital role in differentiating these clusters as $X_2 = 1000$ may be considered as an approximate boundary for the two clusters we estimated. The plot of X_1 versus X_3 seems to cloud the clustering structure as these variables provide almost no information about the estimated two clusters. A 3D scatter plot (see Section 11.4) may further validate this finding.

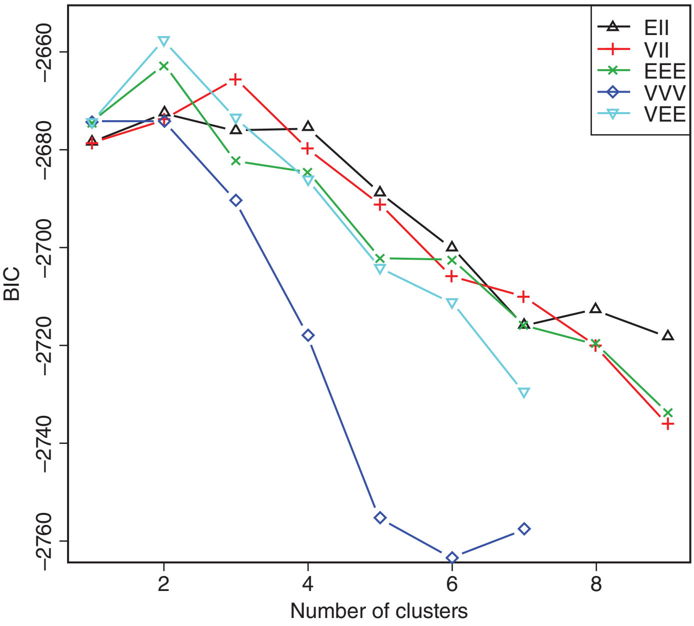
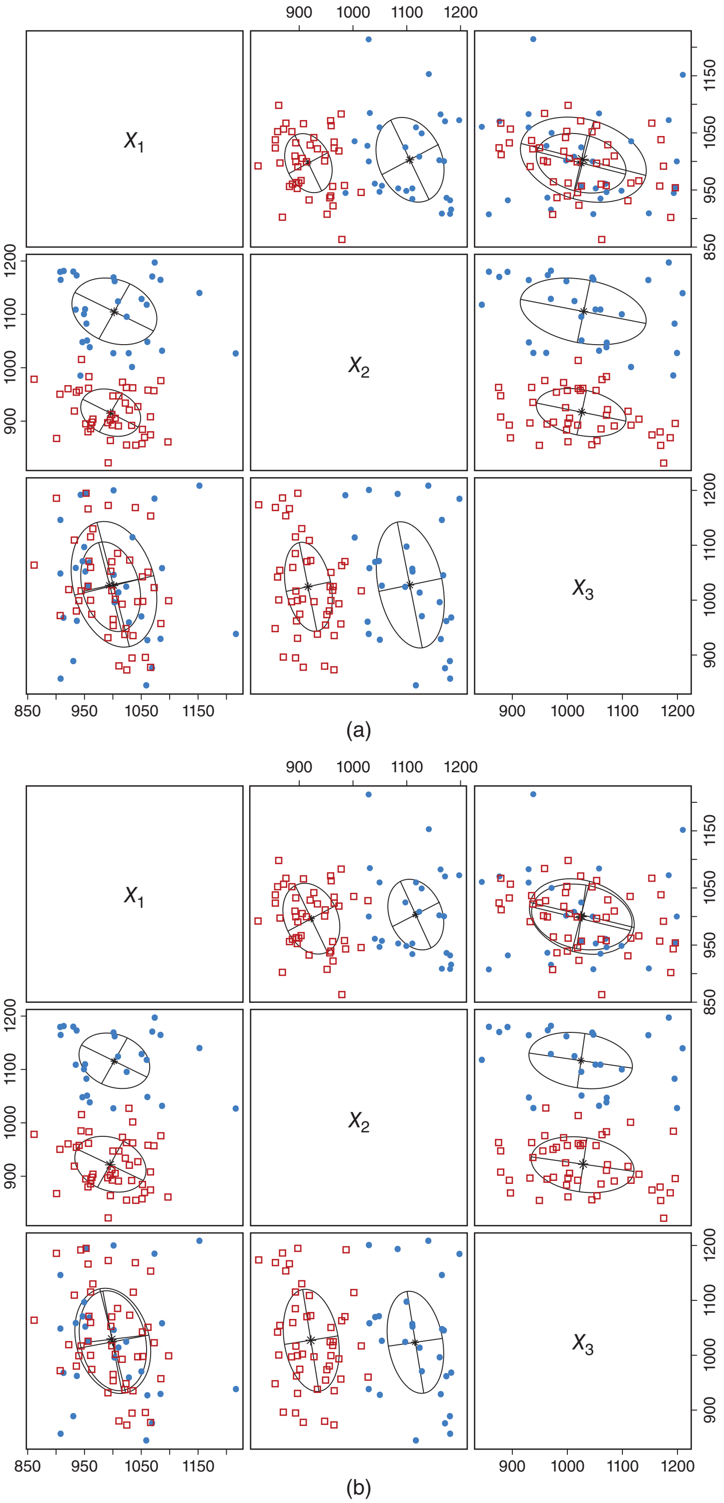


Figure 12.6.1 BIC values for all five covariance structures in Table 12.6.1 for the data of Example 12.6.1.



(a)

(b)

Figure 12.6.2 Estimated clusters based on the covariance structures (a) "VEE" and (b) "EEE" in Table 12.6.1, for the data of Example 12.6.1 in Table 12.6.2.

MINITAB

Model-Based clustering option is not available in MINITAB.

12.7 A CASE STUDY

Case Study (*Seeds Data*)[2] This study focuses on clustering kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian. Researches randomly selected 70 kernels from each variety for the experiment. A nondestructive and considerably cheaper soft X-ray imaging technique that produces high-quality visualization was used to detect the internal kernel structure of wheats. The images were recorded on 13 × 18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The attribute 'target' indicates the wheat variety ($0 =$Kama, $1 =$Rosa, $2 =$Canadian). The data set contains following seven continuous attributes consisting of geometric parameters of wheat kernels. An analyst is interested in clustering these measured attributes that may uniquely characterize wheat variety.

1. $A =$area
2. $P =$perimeter
3. $C =$compactness $= 4 * \pi * A/P^2$
4. $LK =$ length of kernel
5. $WK =$width of kernel
6. $A.Coef =$ asymmetry coefficient
7. $LKG =$ length of kernel groove

The data reported for this case study are available under *case study 12.7.1* on the book website: www.wiley.com/college/gupta/statistics2e.

(a) Use linkage-based clustering methods to cluster these data.
(b) Use Wards' and K-means clustering methods to cluster these data.
(c) Use density-based clustering methods to cluster these data.
(d) Use model-based clustering methods to cluster these data.
(e) Compare your results from parts (a)–(d) with that of attribute 'target'.

[2] **Source:** Charytanowicz et al. (2010). Data available at https://www.kaggle.com/dongeorge/seed-from-uci and more information can be found at https://archive.ics.uci.edu/ml/datasets/seeds.

12.8 USING JMP

This section is not included in the book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Consider the following data set and calculate (a), (b), and (c) below:

Items	X	Y	Z	W
A	13.2	236	58	21.2
B	10.0	263	48	44.5
C	8.1	294	80	31.0
D	8.8	190	50	19.5
E	9.0	276	91	40.6
F	7.9	204	78	38.7
G	3.3	110	77	11.1
H	5.9	238	72	15.8

   (a) Euclidean distance between items A and B?
   (b) The city-block distance between items A and B?
   (c) The cosine correlation coefficient between items A and B?

2. For the data in Review Problem 1,
   (a) Compute the Euclidean distance matrix for all the items.
   (b) Graph the distance matrix using an appropriate graphing tool.
   (c) Describe the main features of the distance matrix.

3. For the data in Review Problem 1,
   (a) Compute the city-block distance matrix for all the items.
   (b) Graph the distance matrix using an appropriate graphing tool.
   (c) Does the city-block distance matrix convey the same information as of the Euclidean distance matrix?

4. Consider the following data set and calculate (a), (b), and (c) below:

A	1	0	1	1	0
B	1	1	0	1	0
C	0	0	1	1	1
D	0	1	0	1	0
E	1	0	1	0	1

   (a) What are the SMCs for all the objects?
   (b) What are the Jaccard coefficients for all the objects?
   (c) Compare your results in part (a) and part (b).

5. Consider the following patient data set. As in Example 12.2.2, introduce appropriate binary variables and calculate the Match-Mismatch coefficient for all the patients. Explain your results.

Patient ID	Gender	Age (yr)	Blood type	Blood pressure (mm Hg)	Glucose level (mg/dl)
P011	F	37	A	120	111
P012	F	61	B	133	122
P013	M	55	A	124	99
P014	M	53	B	112	108
P015	F	58	B	120	109

6. Consider the dichotomous coded variables in Review Problem 5.
   (a) Obtain the city-block distances between all the patients.
   (b) Graph the distance matrix using an appropriate graphing tool.
   (c) How do you compare your results with results from Review Problem 5?

7. Consider the following GPA data set for 10 selected freshman from a certain university. The data consist of students from high school and current college GPA values.

ID	High school GPA	Current college GPA
S1	3.0	3.2
S2	4.0	3.9
S3	2.8	3.1
S4	3.1	3.2
S5	3.7	3.1
S6	3.8	2.9
S7	3.6	3.0
S8	2.9	3.5
S9	3.1	3.9
S10	3.2	3.8

   (a) Calculate the Euclidean distance between all the students with regards to the GPA values.
   (b) Based on part (a), which students seem to be statistically closer and distant than the rest of the students?

8. Refer to the data in Review Problem 7.
   (a) Use the single linkage method to cluster students based on their high school and college GPA values. Report your dendrogram.
   (b) Using part (a), identify the resulting clusters at distance = 0.4.

9. Refer to the data in Review Problem 7.
   (a) Use the complete linkage method to cluster students based on their high school and college GPA values. Report your dendrogram.
   (b) Using part (a), identify the resulting clusters at distance = 0.5.

10. Refer to the data in Review Problem 7.
   (a) Use the average linkage method to cluster students based on their high school and college GPA values. Report your dendrogram.
   (b) Using part (a), identify the resulting clusters at distance = 0.5.

11. Refer to the data in Review Problem 7.
   (a) Use the K-means method to cluster students into four distinct clusters. Report your resulting clusters and their centroids.
   (b) Identify the common features of the resulting clusters in part (a).

12. The K-medoids algorithm can also perform effective clustering. That is, in the K-means algorithm, we replace mean centroid with median centroid and perform rest of the calculations accordingly. Illustrate the strength and weakness of the K-means in comparison with the K-medoids algorithm (see Section 12.4.1).

13. Refer to the data in Review Problem 7 and the K-mediods discussion in Review Problem 12.
   (a) Use the K-mediods method to cluster students into four distinct clusters using the city-block distance. Report your cluster results. (*Hint*: you may use ‘kmed’ library in R for this purpose).
   (b) Compare your results in (a) with that of the K-means results.

14. Refer to the data in Review Problem 7. Use the Density-Based method to cluster students into distinct clusters (*Hint*: use eps = 0.3 and minPts = 2 in your DBSCAN algorithm).

15. Consider the following data set.

Object	A	B	C	D	E	F	G	H	I	J
X	0	1	2	2	4	5	6	6	6	5
Y	1	0	0	6	8	6	2	1	1	8

   (a) Use the K-means algorithm to split the above 10 objects into three distinct clusters by considering objects A, B, and C as their initial cluster centroids.
   (b) Using part (a), plot the resulting clusters after each iteration along with their centroids.

16. Consider the dichotomous coded variables in Review Problem 5.
   (a) Use the single, complete, and the average linkage methods to construct the dendrograms.
   (b) Compare your results in part (a).

17. The following data contains US arrest data published in McNeil (1977). The table shows arrests per 100,000 residents for "assault," "murder," and "rape" in each of the 50 US states in 1973. The variable "UP" indicates the percent of the urban population living in urban areas.

State	Murder	Assault	UP	Rape	State	Murder	Assualt	UP	Rape
Alabama	13.2	236	58	21.2	Montana	6.0	109	53	16.4
Alaska	10.0	263	48	44.5	Nebraska	4.3	102	62	16.5
Arizona	8.1	294	80	31.0	Nevada	12.2	252	81	46.0
Arkansas	8.8	190	50	19.5	New Hampshire	2.1	57	56	9.5
California	9.0	276	91	40.6	New Jersey	7.4	159	89	18.8
Colorado	7.9	204	78	38.7	New Mexico	11.4	285	70	32.1
Connecticut	3.3	110	77	11.1	New York	11.1	254	86	26.1
Delaware	5.9	238	72	15.8	North Carolina	13.0	337	45	16.1
Florida	15.4	335	80	31.9	North Dakota	0.8	45	44	7.3
Georgia	17.4	211	60	25.8	Ohio	7.3	120	75	21.4
Hawaii	5.3	46	83	20.2	Oklahoma	6.6	151	68	20.0
Idaho	2.6	120	54	14.2	Oregon	4.9	159	67	29.3
Illinois	10.4	249	83	24.0	Pennsylvania	6.3	106	72	14.9
Indiana	7.2	113	65	21.0	Rhode Island	3.4	174	87	8.3
Iowa	2.2	56	57	11.3	South Carolina	14.4	279	48	22.5
Kansas	6.0	115	66	18.0	South Dakota	3.8	86	45	12.8
Kentucky	9.7	109	52	16.3	Tennessee	13.2	188	59	26.9
Louisiana	15.4	249	66	22.2	Texas	12.7	201	80	25.5
Maine	2.1	83	51	7.8	Utah	3.2	120	80	22.9
Maryland	11.3	300	67	27.8	Vermont	2.2	48	32	11.2
Massachusetts	4.4	149	85	16.3	Virginia	8.5	156	63	20.7
Michigan	12.1	255	74	35.1	Washington	4.0	145	73	26.2
Minnesota	2.7	72	66	14.9	West Virginia	5.7	81	39	9.3
Mississippi	16.1	259	44	17.1	Wisconsin	2.6	53	66	10.8
Missouri	9.0	178	70	28.2	Wyoming	6.8	161	60	15.6

(a) Use this data to cluster US states using the single, complete, and the average linkage methods.

(b) Use this data to cluster US states using the Ward's method.

18. Refer to the data in Review Problem 17. Use the K-means method to cluster US states into six distinct clusters.

19. Refer to the data in Review Problem 17. Use the Density-Based method to cluster US states (*Hint*: use eps = 20 and minPts = 3 in your DBSCAN algorithm).

20. Refer to the data in Review Problem 17.
    (a) Use the BIC option in 'mclust' package in R to select the appropriate number of model-based clusters and the covariance structure for this data. Note that by the default 'Mclust()' function produces 14 different models.
    (b) Explain basic features of the structure you selected in part (a).
    (c) Plot resulting clusters against the crime types and explain any interesting trends you see in those clusters.

21. Refer to the data in "Review Problem 21" on the website: www.wiley.com/college/gupta/statistics2e.
    (a) Using the covariance structures listed in Table 12.6.1 select the appropriate number of clusters for this data.
    (b) Based on the BIC values which covariance structure seems reasonable for this data?

Chapter 13

ANALYSIS OF CATEGORICAL DATA

The focus of this chapter is on the development of chi-square goodness-of-fit tests used as nonparametric procedures.

Topics Covered

- Chi-square goodness of fit tests to determine if the sample data come from some specified probability model.
- The chi-square test of a hypothesis that the two factors cross-classifying a sample (count or frequency) data are independent.
- Use of 2×2 and $r \times s$ contingency tables to test a hypothesis that the populations under investigation are homogeneous with respect to certain criteria.

Learning Outcomes

After studying this chapter, the reader will be able to

- Use the chi-square goodness of fit test to evaluate certain distributional assumptions.
- Test whether or not two classifications of a population are independent.
- Use contingency tables to test whether populations are homogeneous with respect to some characteristics of interest.

13.1 INTRODUCTION

Often data collected by an investigator through experimentation, observation, or a sample survey are classified into various categories, and frequency counts of observations in

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

each category are recorded. For example, a manager of a manufacturing company may be interested in finding the number of variously sized rods available in stock or the number of defective parts produced during different work shifts. A sociologist may be interested in finding the number of persons of different religious faiths, different political party affiliations, different races, or different income groups within a large metropolitan area. Sometimes quantitative data can also be classified into different categories so that the observed data are also categorical data. For example, a person may be classified as overweight, normal weight, or underweight, or a person's number of years of schooling may be classified as under 10 years, 10–12 years, 12–16 years, or over 16 years. In this chapter, we discuss the chi-square tests used to analyze categorical data.

13.2 THE CHI-SQUARE GOODNESS-OF-FIT TEST

Goodness-of-fit tests arise when testing certain hypotheses on the basis of a sample that consists of n independent observations from a given population. The measurement scale of these observations is at least of a *nominal* type (see Chapter 2). For example, the observations may be classified into k categories, so that each observation belongs to one and only one category and no observation is left out. Then, the data can be presented in the form of a table (see Table 13.2.1) consisting of k cells in which each cell corresponds to one of the k categories. The number of observations in each cell is called the *observed cell frequency.* This scenario can also be looked upon as an experiment consisting of n independent trials such that each trial has k possible outcomes and each observation represents an outcome of a trial.

Table 13.2.1 Results of n observations classified into k categories.

Outcomes (categories)	A_1	A_2	...	A_k
Observed frequency	f_1	f_2	...	f_k
Theoretical frequency	$(n\theta_1)$	$(n\theta_2)$	...	$(n\theta_k)$

Suppose that we want to compare the observed cell frequencies with their expected (or theoretical) frequencies. In general, we let $A_1, A_2, \ldots, A_k$ be k mutually exclusive and exhaustive outcomes or categories, and let the probabilities of these outcomes (i.e. the probabilities of an observation belonging to categories $A_i, i = 1, 2, \ldots, k$) be $\theta_1, \theta_2, \ldots, \theta_k$, respectively, where $\theta_1, \theta_2, \ldots, \theta_k$ are all positive and $\theta_1 + \cdots + \theta_k = 1$.

Suppose that n independent trials of the experiment are made, and let $f_1, f_2, \ldots, f_k$ be the number of trials that result in outcomes $A_1, A_2, \ldots, A_k$, respectively. The numbers $n\theta_1, n\theta_2, \ldots, n\theta_k$ in parentheses are the expected number of trials that result in outcomes $A_1, A_2, \ldots, A_k$, respectively. Then $f_1, f_2, \ldots, f_k$ are random variables with the multinomial distribution (see Section 4.9)

$$\frac{n!}{f_1!f_2!\ldots!f_k!}\theta_1^{f_1}\theta_2^{f_2}\cdots\theta_k^{f_k} \tag{13.2.1}$$

where $\sum_{i=1}^{k} f_i = n$. The expectation of f_i is $E(f_i) = n\theta_i, \quad i = 1, 2, \ldots, k$.

If we want to test the null hypothesis H_0 that our sample of size n comes from a multinomial population with probabilities $P(A_i) = \theta_i$, $i = 1, 2, \ldots, k$, where $\theta_1, \theta_2, \ldots, \theta_k$ are known, we use the *chi-square* test statistic, denoted by χ^2 and defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\theta_i)^2}{n\theta_i} \tag{13.2.2}$$

If the observed values of f_i are all exactly equal to $n\theta_i$, we have a "perfect fit" and $\chi^2 = 0$. Thus, large values of χ^2 will tend to discredit the null hypothesis, and smaller values of χ^2 will tend to confirm the hypothesis that the observed frequencies f_i are not significantly different from the expected frequencies $n\theta_i$, $i = 1, 2, \ldots, k$.

For *moderately large values of* n, the distribution of the test statistic given in (13.2.2) is approximately the chi-square distribution having $k - 1$ degrees of freedom. The degrees of freedom used are $k - 1$, not k. This is because the observed frequencies f_i, $1 = 1, 2, \ldots, k$, must obey one constraint (or condition), namely $\sum_{i=1}^{k} f_i = n$ the sample size. Thus, the multinomial model (13.2.1) is a probability function of $k - 1$ variables, say $f_1, \ldots, f_{k-1}$, with f_k given by $f_k = n - \sum_{i=1}^{k-1} f_i$. In practice, we usually require that the expected frequency $n\theta_i \geq 5$ for each i. (The reader should be aware that there is no one opinion about the minimum number of expected frequencies. For example the minimum expected frequency in each cell proposed by Cochran, (1952), is 1.) If there is difficulty in meeting this criterion of minimum expected frequencies, we may combine two or more of the outcomes, say $A_{i_1}, \ldots, A_{i_r}$, with the smallest probabilities, say $\theta_{i_1}, \ldots, \theta_{i_r}$, into a single outcome until the condition $n(\theta_{i_1} + \cdots + \theta_{i_r}) \geq 5$ is met. We illustrate this procedure in Examples 13.2.1 and 13.2.2.

If we test the hypothesis at the α level of significance, the observed value of χ^2 is considered to be significant at this level if $\chi^2 > \chi^2_{k-1;\,\alpha}$; that is if we have that $\chi^2 > \chi^2_{k-1;\alpha}$, we reject the hypothesis that the observations come from a distribution for which $P(A_i) = \theta_i$, i=1, 2, $\ldots$, k. Alternatively, we may use the p-value, $P(\chi^2_{k-1} > obs\ \chi^2)$, and if the p-value is less than or equal to α, we reject the hypothesis.

If the $\theta_1, \theta_2, \ldots, \theta_k$ are unknown but expressible in terms of c parameters that have to be estimated from $f_1, \ldots, f_k$, then the resulting χ^2 statistic has approximately the chi-square distribution with $(k - c - 1)$ degrees of freedom. Of course, in this case, the observed value of χ^2 is considered to be significant if it exceeds $\chi^2_{k-c-1;\,\alpha}$.

Example 13.2.1 (Icosahedral die) *An icosahedral die has two sides marked 1, two sides marked 2, …, and two sides marked 0, which will be designated as 10, so an icosahedral die has 20 faces with each digit on two faces. Test whether the die is behaving as a "true" or "fair" die on the basis of the 200 throws whose outcomes are tabulated in Table 13.2.2.*

Solution: If the die is true, then the probability of obtaining any of the numbers 1–10 in a single toss is 2/20 = 1/10, and we would then expect that each number turns up $200 \times 1/10 = 20$ times in 200 throws. The question we ask, then, is whether the set of observed f_i is compatible with the null hypothesis that the die is true, that is whether $\theta_i = 1/10$ for each i. We note that $n\theta_i \geq 5$ for each i and apply the χ^2 test. In this example,

Table 13.2.2 Results and analysis of 200 throws of an icosahedral die.

x_i	f_i	$n\theta_i$	$(f_i - n\theta_i)^2$	$\dfrac{(f_i - n\theta_i)^2}{n\theta_i}$
1	17	20	9	0.45
2	19	20	1	0.05
3	26	20	36	1.80
4	18	20	4	0.20
5	16	20	16	0.80
6	23	20	9	0.45
7	21	20	1	0.05
8	24	20	16	0.80
9	20	20	0	0.00
10	16	20	16	0.80
	200	200		5.40

$k = 10$ and, from Table 13.2.2, we obtain

$$\sum_{i=1}^{10} \frac{(f_i - n\theta_i)^2}{n\theta_i} = 5.40$$

Consulting Table A.4 and choosing $\alpha = 0.05$, we find that $\chi^2_{9;.05} = 16.92$. Since the observed value of χ^2 is 5.40, which is less than 16.92, we do not reject the hypothesis that the die is true, accepting $\theta_i = 1/10$ for each $i = 1, \ldots, 10$; to put it another way, we can say that the results of the 200 throws do not contradict the hypothesis of a true die at the 5% level of significance.

Example 13.2.2 (Alpha particles emitted from uranium) *It is believed that when a certain type of uranium is placed in a radioactive counter for a given interval of time, the number* X *of α-particles emitted during the interval behaves like a random variable having the Poisson distribution, that is,*

$$P(X = x) = p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots \tag{13.2.3}$$

where λ is an unknown parameter.

In an experiment, the number of emissions from a piece of uranium is determined for each of 100 time intervals of equal length, with the results shown in Table 13.2.3. We want to test a hypothesis that the data behave as a sample from a population having the Poisson distribution.

Solution: Since λ is unknown, it must be estimated. With the use of the method of maximum likelihood, it is found that the "best" estimator for λ is the sample average $\bar{X}$ (see Section 8.2). From Table 13.2.3, we find $\bar{X} = 4.2$ (the average number of emissions per time interval). Hence, if X has a Poisson distribution, we estimate $p(x)$ as follows:

$$p(x) = (e^{-4.2}(4.2)^x)/x!$$

Table 13.2.3 Results of $\boldsymbol{\alpha}$ emission counting experiment on uranium.

$\boldsymbol{\alpha}$-Particles emitted, x_i	Observed time intervals, f_i	Expected time intervals	$x_i f_i$
0	1	1.5	0
1	5	6.3	5
2	16	13.2	32
3	17	18.5	51
4	26	19.4	104
5	11	16.3	55
6	9	11.4	54
7	9	6.9	63
8	2	3.6	16
9	1	1.7	9
10	2	0.7	20
11	1	0.4	11
	100	99.9	420

Table 13.2.4 Results of Table 13.2.3 after grouping some categories.

Observed time intervals, f_i	Number (estimated) expected, $n\theta_i$	$(f_i - n\theta_i)^2$	$\dfrac{(f_i - n\theta_i)^2}{n\theta_i}$
6	7.8	3.24	0.415
16	13.2	7.84	0.594
17	18.5	2.25	0.122
26	19.4	43.56	2.245
11	16.3	28.09	1.723
9	11.4	5.76	0.505
9	6.9	4.41	0.639
6	6.4	0.16	0.025
			6.268

We estimate that $np(x) = 100(e^{-4.2}(4.2)^x/x!)$ time intervals will emit X α-particles, where $x = 0, 1, 2, \ldots,$ as shown in column 3 of Table 13.2.3. Note that several of the expected frequencies are less than 5. We proceed by grouping the *adjacent* classes until the estimated "expected frequencies" are greater than or equal to 5. For example, we group the first with the second, group the last four, and apply the χ^2 test to the resulting data, as seen in Table 13.2.4.

As shown in Table 13.2.4, the chi-square test is applied to $k = 8$ classes. Furthermore, the expected frequencies are calculated by using estimates of $n\theta_1, n\theta_2, \ldots, n\theta_8$, where each θ_i is computed from the Poisson distribution (13.2.3) after estimating the single parameter λ. Hence, the number of degrees of freedom involved in applying the chi-square test is $(8 - 1 - 1) = 6$ because $k = 8$, $c = 1$. In fact, there are two constraints on the f_i, that is, $\sum f_i = n$ and $\sum f_i x_i/n = \hat{\lambda}$, accounting for the loss of $c + 1 = 1 + 1 = 2$ degrees of freedom, so that $8 - 2 = 6$ is used for the degrees of freedom.

The upper 5% significance point of χ_6^2 is 12.59 and the observed $\chi_6^2 = 6.268$. Therefore, we do not reject the hypothesis that the data behave as a sample coming from a Poisson distribution, and estimating the mean of the distribution as 4.2.

We illustrate the use of MINITAB with the following example.

Example 13.2.3 (Patients admitted in a hospital) *The following data show the number of patients admitted in a hospital during intervals of one-hour over a period of five days* ($5 \times 24 = 120$ *one-hour intervals).*

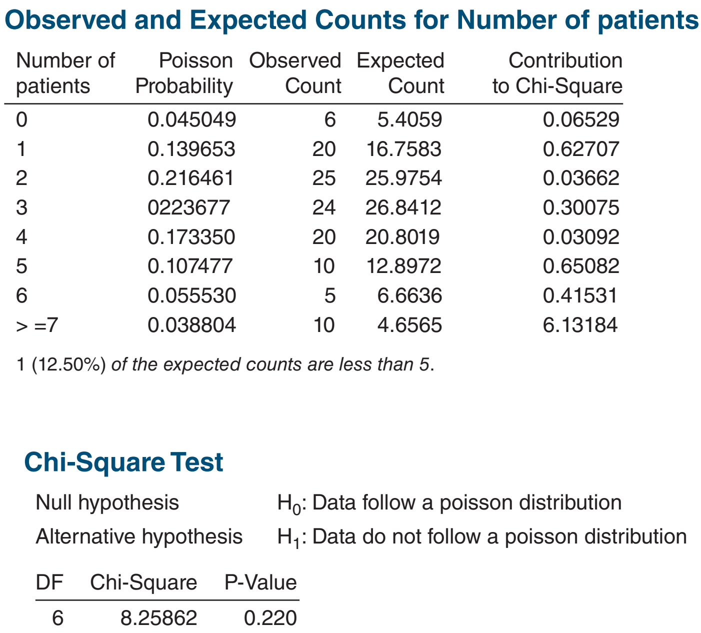
Number of patients:	0	1	2	3	4	5	6	7 or more
Frequency:	6	20	25	24	20	10	5	10

Test at the 5% level of significance that the data come from the Poisson distribution.

Solution:

MINITAB

1. Enter the data in columns C1 (number of patients) and C2 (frequency).
2. From the Menu bar select **Stat** > **Basic Statistics** > **Goodness-of-fit Test for Poisson**.
3. Enter C1 in the box next to **Variables** and C2 in the box next to **Frequency variable**.
4. Select any desired options: **Graph** and **Results**, and click **OK**. The following MINITAB output appears in the Session window.



```
Observed and Expected Counts for Number of patients

Number of    Poisson  Observed  Expected      Contribution
patients  Probability    Count     Count   to Chi-Square
0            0.045049        6    5.4059         0.06529
1            0.139653       20   16.7583         0.62707
2            0.216461       25   25.9754         0.03662
3             0223677       24   26.8412         0.30075
4            0.173350       20   20.8019         0.03092
5            0.107477       10   12.8972         0.65082
6            0.055530        5    6.6636         0.41531
> =7         0.038804       10    4.6565         6.13184

1 (12.50%) of the expected counts are less than 5.


Chi-Square Test

Null hypothesis          H0: Data follow a poisson distribution
Alternative hypothesis   H1: Data do not follow a poisson distribution

DF  Chi-Square  P-Value
 6     8.25862    0.220
```

Since the p-value is 0.22, (see the above MINITAB output) which is greater than the 0.05 the level of significance, we do not reject the null hypothesis. (*Note*: The p-value here is $P(\chi_6^2 > 8.25862) = 0.220$.)

USING R

The following manual R code can be used to conduct the chi-squared goodness of fit test for Poisson data.

```
#Assign data
x = c(0, 1, 2, 3, 4, 5, 6 , 7)
f = c(6, 20, 25, 24, 20, 10, 5, 10)

#Estimate λ̂
lambda = sum(f*x)/sum(f)

#Calculate expected counts
expected = sum(f)*dpois(x[1:7],lambda)
expected[8] = sum(f)*(1-ppois(x[7],lambda))

#chi.squared contributions
chi.squared = (f - expected)^2/expected

#Observed and Expected counts and contribution to chi-square
goodness.of.fit = cbind(x, f, expected)
colnames(goodness.of.fit) = c(‘No. of Patients’, ‘Frequency’, ‘Expected’, ‘chi.squared’)
goodness.of.fit # This will produce a table as in above MINITAB Step 4.

#Test statistic and p-value
chi.squared.statistic = sum(chi.squared)
chi.squared.statistic #R output
[1] 8.258618

p. value = pchisq(chi.squared.statistic, length(f)-2, lower.tail = F)
p. value #R output
[1] 0.219762
```

Example 13.2.4 (Diastolic blood pressures) *The diastolic blood pressure of a random sample of 100 male adults between the age of 30 and 35 years is measured. The data are presented in the form of a frequency distribution table in Table 13.2.5. We want to verify that these data come from a normal population with mean μ and variance σ^2. Use $\alpha = 0.01$.*

Solution: Here we test the hypothesis:

H_0: The sample comes from a normal population
H_1: The sample does not come from a normal population

We apply the chi-square test (13.2.2) to examine the hypothesis that the frequencies in Table 13.2.5 behave as if they come from a normal distribution.

Since the mean μ and variance σ are unknown, they must be estimated. Using the results of Section 2.7 for the mean and the variance of grouped data, we have

$$\bar{X} = 83.10, \quad S = 5.698$$

Table 13.2.5 Frequency distribution of diastolic blood pressures of 100 males.

Class	Class *z*-values	Observed frequencies (f_i)	Probabilities (θ_i)	Expected frequencies $n\ \theta_i$
$(-\infty - 70)$	<-2.30	0	0.0107	1.07
$[70 - 75)$	$-2.30, -1.42$	10	0.0671	6.71
$[75 - 80)$	$-1.42, -0.54$	15	0.2168	21.68
$[80 - 85)$	$-0.54,\ 0.33$	40	0.3347	33.47
$[85 - 90)$	$0.33,\ 1.21$	25	0.2576	25.76
$[90 - 95)$	$1.21,\ 2.09$	8	0.0948	9.48
$[95 - 100)$	$2.09,\ 2.97$	2	0.0172	1.72
$[100 - \infty)$	> 2.97	0	0.0011	0.11
Total		100	1.00	100

which we use as estimates of population mean μ and standard deviation σ, respectively.

Next, we test the hypothesis H_0, which says that the sample comes from a normal population. We approximate the normal distribution by the distribution $N(\bar{X},\ S^2)$. In this example n is large ($n = 100$), so we use the $N(83.10,\ (5.698)^2)$ distribution. We then calculate $z =$ (Class boundary $-$ 83.10)/5.698 for the class *z*-values. For example (see columns 1 and 2 of Table 13.2.5), a class boundary 80 has *z*-value $(80 - 83.10)/5.698 = -0.54$.

Hence, to estimate the probability θ_i of an observation falling in a class, we proceed by letting X be distributed as $N(83.10,\ (5.698)^2)$. For example $\hat{\theta}_3 = P(X$ falls in the class [75-80)), is given by

$$P(75 \leq X < 80) = P\left(\frac{75 - 83.10}{5.698} \leq Z \leq \frac{80 - 83.10}{5.698}\right) = P(-1.42 \leq Z \leq -0.54)$$

or $\hat{\theta}_3 = \Phi(-0.54) - \Phi(-1.42) = 0.2946 - 0.0778 = 0.2168$. Thus we estimate the expected frequency of the third class to be approximately $n\hat{\theta}_3 = 100(0.2168) = 21.68$.

We note that we have imposed three constraints on the sample data: $\sum f_i = n$, and estimating the population mean and standard deviation by $\sum f_i X_i / n = \bar{X}$, $\sqrt{\sum f_i (X_i - \bar{X})^2/(n-1)} = S$, respectively, so we lose three degrees of freedom, ultimately accounting for the degrees of freedom used in Table 13.2.6. Columns 4 and 5 of Table 13.2.5 were obtained under the null hypothesis of normality. Since several of the expected frequencies in Table 13.2.5 are less than 5, we proceed by grouping the classes until all the "expected frequencies" are greater than or equal to 5. This grouping gives rise to the frequency table in Table 13.2.6.

In the present table, Table 13.2.6, $k = 5$ and $c = 2$ since both μ and σ are estimated, and hence, as previously observed, the number of degrees of freedom is $5 - 2 - 1 = 2$. The observed value of χ^2 is 4.1398, which is less than the upper 5% point $\chi^2_{2,.05} = 5.991$. Therefore, at the 5% level of significance, we do not reject the hypothesis that the sample,

Table 13.2.6 Results after grouping in Table 13.2.5.

Observed frequency f_i	Expected frequency $n\theta_i$	$\dfrac{(f_i - n\theta_i)^2}{n\theta_i}$
10	7.78	0.6335
15	21.68	2.0582
40	33.47	1.2740
25	25.76	0.0224
10	11.31	0.1517
Total		4.1398

which produced the frequencies in column 3 of Table 13.2.5, behaves as frequencies coming from a normal distribution.

PRACTICE PROBLEMS FOR SECTION 13.2

1. The table below gives the month of birth of a sample of 756 artists. Test, at the 5% level of significance, the null hypothesis that there is no seasonal variation in the months of the year in which artists are born.

Birth month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Frequency	68	78	67	60	61	51	50	60	67	61	73	60	756

2. An engineering society is interested in finding if males and females have the same interest in graduate work in engineering. The society surveyed 100 graduate programs in engineering each of which had admitted seven PhD students. The data below give the distribution of males and females among those students who were admitted to the 100 engineering programs. Test at the 5% level of significance the hypothesis that the males and females have the same interest in graduate work in engineering.

Number of males	7	6	5	4	3	2	1	0
Number of females	0	1	2	3	4	5	6	7
Number of programs	24	20	16	12	10	8	8	2

3. The Occupational Safety and Health Administration revealed the data below on safety violations per week by a manufacturing company over a period of 80 weeks. Fit a Poisson distribution to these data and test the null hypothesis at the 1% level of significance that the data behave like a sample from a Poisson population.

Number of violations	0	1	2	3	4	5
Number of weeks	25	15	8	14	7	11

4. Fit a normal distribution to the data of Problem 22 in Review Practice Problems of Chapter 2 and test the null hypothesis at 5% level of significance that the data behave like a sample from a normal population. The data of Problem 22 is reproduced below. Find the observed level of significance.

Age	[35-40)	[40-45)	[45-50)	[50-55)	[55-60)	[60-65)
Frequency	60	75	68	72	90	55

5. It is believed that when a certain type of uranium is placed in a radioactive counter for a given interval of time, the number X of gamma particles emitted during the interval behaves as a random variable having the Poisson distribution. In an experiment, the number of emissions from a piece of uranium was obtained for each of 50 intervals of equal length with the results shown below. Use the chi-square goodness-of-fit test to test the null hypothesis at the 5% level of significance that the data behave like a sample from a Poisson population.

Number of gamma particles	0	1	2	3	4	5	6
Observed time intervals	2	3	10	12	14	5	4

6. In 100 throws of a single die, the data obtained are shown below. Test at the 5% level of significance the null hypothesis that the die is a fair die.

Number of points	1	2	3	4	5	6
Number of throws	10	25	20	16	19	10

7. It is believed that the number of cars passing through a toll booth in a given interval of time is a random variable having Poisson distribution with unknown parameter λ. In an experiment, the number of cars passing through the toll booth is determined for each of 100 time intervals of equal length, with the results shown below. Test the hypothesis that the data behave like a sample from a population having the Poisson distribution. Use $\alpha = 0.05$.

Number of cars	0	1	2	3	4	5	6	7	8	9	10
Frequency	5	7	8	10	12	10	8	11	14	10	5

8. The following data give the grades in a required engineering course for a given semester. Test at the 5% level of significance that the data follow a multinomial distribution, where $\theta_1 = 0.22,\ \theta_2 = 0.28, \theta_3 = 0.30,\ \theta_4 = 0.08,\ \theta_5 = 0.07,\ \theta_6 = 0.05$. Use $\alpha = 0.05$.

Grade	A	B	C	D	F	Incomplete
Frequency	20	30	25	10	12	3

9. The following data give the scores made by a basketball player in 40 consecutive games. Test at the 1% level of significance that these data follow a normal distribution.

34	36	34	34	36	30	31	35	31	36	34	35	36	25	36	25	30	33	27	36
31	26	24	25	33	36	33	34	35	26	25	33	29	27	33	31	33	24	35	25

10. The data below give the number of flights arriving late at a regional airport during the last 10 weeks. Test the hypothesis, at the 5% level of significance, that the number of flights arriving late is the same for each of the past 10 weeks.

Week	1	2	3	4	5	6	7	8	9	10
Frequency	12	8	14	16	8	7	12	12	16	15

13.3 CONTINGENCY TABLES

13.3.1 The 2 × 2 Case with Known Parameters

Suppose that the outcome of each trial in an experiment can be classified into one and only one of the four mutually exclusive and exhaustive classes $A \cap B$, $A \cap \bar{B}$, $\bar{A} \cap B$, $\bar{A} \cap \bar{B}$ with probabilities $\theta_1\tau_1, \theta_1\tau_2, \theta_2\tau_1, \theta_2\tau_2$, respectively, where θ_1, θ_2, τ_1, τ_2 are all positive, $\theta_1 + \theta_2 = 1$, and $\tau_1 + \tau_2 = 1$. This means that factors A and B are independent and that $P(A) = \theta_1, P(\bar{A}) = \theta_2, P(B) = \tau_1, P(\bar{B}) = \tau_2$. These four classes, $A \cap B$, $A \cap \bar{B}$, $\bar{A} \cap B$, $\bar{A} \cap \bar{B}$, and their probabilities under independence of A and B can be arranged as shown in Table 13.3.1.

Table 13.3.1 Classes and their probabilities in a 2 × 2 contingency table in which rows and columns are independent.

	B	$\bar{B}$	
A	$\theta_1\tau_1$	$\theta_1\tau_2$	θ_1
$\bar{A}$	$\theta_2\tau_1$	$\theta_2\tau_2$	θ_2
	τ_1	τ_2	1

Now suppose that n independent trials are performed and that f_{11} trials result in class $A \cap B$, f_{12} in $A \cap \bar{B}$, f_{21} in $\bar{A} \cap B$, and f_{22} in $\bar{A} \cap \bar{B}$ so that $f_{11} + f_{12} + f_{21} + f_{22} = n$. The experimental results can be arranged as shown in Table 13.3.2, where the marginal totals $f_{1\cdot}, f_{2\cdot}, f_{\cdot 1}, f_{\cdot 2}$ are defined as follows:

$$f_{1\cdot} = f_{11} + f_{12}, \quad f_{2\cdot} = f_{21} + f_{22}, \quad f_{\cdot 1} = f_{11} + f_{21}, \quad f_{\cdot 2} = f_{12} + f_{22}$$

Note that $f_{1\cdot} + f_{2\cdot} = n$, $f_{\cdot 1} + f_{\cdot 2} = n$.

Table 13.3.2 Classes and their frequencies in n independent trials of a 2×2 experiment.

	B	$\bar{B}$	
A	f_{11}	f_{12}	$f_{1\cdot}$
$\bar{A}$	f_{21}	f_{22}	$f_{2\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	n

Under the assumption of independence of the A and B factors, the frequencies $f_{11}, f_{12}, f_{21}, f_{22}$ are random variables having the multinomial distribution

$$\frac{n!}{f_{11}!f_{12}!f_{21}!f_{22}!}(\theta_1\tau_1)^{f_{11}}(\theta_1\tau_2)^{f_{12}}(\theta_2\tau_1)^{f_{21}}(\theta_2\tau_2)^{f_{22}} \qquad (13.3.1)$$

where $0 \leq f_{ij} \leq n$ and $\sum_{i=1}^{2} \sum_{j=1}^{2} f_{ij} = n$.

It is easy to see that under the assumption of the independence of factors A and B, any one of the f_{ij} has the binomial distribution of size n and parameter $\theta_i\tau_j$. It then follows that $\mathrm{E}(f_{ij}) = n\theta_i\tau_j$, or n times the entries of Table 13.3.1.

If we know the values of $\theta_1, \theta_2, \tau_1, \tau_2$, we can test the assumption of independence of the A and B factors using the test statistic given below in (13.3.2). We note that $k = 4$ (the four categories are $A \cap B$, $A \cap \bar{B}$, $\bar{A} \cap B$, and $\bar{A} \cap \bar{B}$), and that there is one constraint on the f_{ij}'s, namely $\sum_i \sum_j f_{ij} = n$.

Further, in the discussion above, the experiment factors A and B may be looked upon as two criteria cross-classifying the observations in a sample so that each observation in the sample pertains to one and only one category of each criterion. For example, the two criteria may be income (A) and gender (B), which may be defined as $A \leq 50{,}000$, $\bar{A} > 50{,}000$ with B a male and $\bar{B}$ a female. Thus, we want to test the following hypothesis:

H_0 : The two criteria of cross-classifications are independent H_1 : The two criteria of cross-classifications are not independent

When the null hypothesis is true, we have the following result:

Theorem 13.3.1 *If $f_{11}, f_{12}, f_{21}, f_{22}$ are the observed frequencies shown in Table 13.3.2, then for large n, the quantity*

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(f_{ij} - n\theta_i\tau_j)^2}{n\theta_i\tau_j} \qquad (13.3.2)$$

has, approximately, the chi-square distribution with three degrees of freedom.

Hence, if the observed value of χ^2 determined by (13.3.2) exceeds $\chi^2_{3,\alpha}$, we would reject the hypothesis that the A and B classifications are independent at the $100\alpha\%$ level

Table 13.3.3 Data on two strains of *Drosophila* ($n = 600$).

	$B(2/3)$	$\bar{B}(1/3)$
$A(3/4)$	(300) 313	(150) 135
$\bar{A}(1/4)$	(100) 93	(50) 59

of significance, and the sample evidence, then, would support the assertion that there is a significant degree of dependence between the A and B classifications at the $100\alpha\%$ level of significance.

Example 13.3.1 (Cross breeding) *A cross-breeding experiment with two strains of* Drosophila *is conducted to determine whether or not eye color, say* A*, and wing type, say* B*, are independent characteristics. It is known that the probability of progeny of the strains of* Drosophila *having dull-colored eyes* A *is 3/4, while the probability of progeny of these two strains having type* B *wing is 2/3. Six hundred progeny (n = 600) are selected at random, and each are classified as to their eye color and wing type. The resulting frequencies f_{ij} are given in Table 13.3.3. The expected number for each of the four categories, that is, E(f_{ij}), under the assumption of independence, are the entries in Table 13.3.3 in parentheses. Does the sample evidence support the assumption of independence at the 5% level?*

Solution: The observed value of

$$\chi_3^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(f_{ij} - n\theta_i\tau_j)^2}{n\theta_i\tau_j}$$

with $n = 600$; $\theta_1 = 3/4$, $\theta_2 = 1/4$, $\tau_1 = 2/3$, $\tau_2 = 1/3$, is given by

$$\begin{aligned}\chi_3^2 &= \frac{(313-300)^2}{300} + \frac{(135-150)^2}{150} + \frac{(93-100)^2}{100} + \frac{(59-50)^2}{50}\\ &= 0.56333 + 1.50000 + 0.49000 + 1.62000\\ &= 4.17333\end{aligned}$$

The observed value of χ^2 is 4.17333, which is less than the upper 5% point $\chi^2_{3;0.05} = 7.8147$. Therefore, at the 5% level of significance, we do not reject the hypothesis of independence of the factors of eye color and type of wing.

13.3.2 The 2 × 2 Case with Unknown Parameters

Consider, again, the experimental situation of Section 13.3.1 but with θ_1, θ_2, τ_1, τ_2 not known. These values must be estimated from the experimental results summarized in Table 13.3.2. Again, it is easy to see that any one of the so-called marginal totals $f_{1.}$ or $f_{2.}$ or $f_{.1}$ or $f_{.2}$ have the binomial distribution. For example $f_{1.}$ is the number of items in the sample that has the attribute A, so $f_{1\cdot}$ has expectation $n\theta_1$ and $\hat{\theta}_1 = f_{1\cdot}/n$ is an unbiased point estimator of θ_1.

Summarizing, we have that

$$\begin{aligned}
\hat{\theta}_1 &= \text{estimator for} \quad \theta_1 = f_{1\cdot}/n, \quad f_{1\cdot} = f_{11} + f_{12} \\
\hat{\theta}_2 &= \text{estimator for} \quad \theta_2 = f_{2\cdot}/n, \quad f_{2\cdot} = f_{21} + f_{22} \\
\hat{\tau}_1 &= \text{estimator for} \quad \tau_1 = f_{\cdot 1}/n, \quad f_{\cdot 1} = f_{11} + f_{21} \\
\hat{\tau}_2 &= \text{estimator for} \quad \tau_2 = f_{\cdot 2}/n, \quad f_{\cdot 2} = f_{12} + f_{22}
\end{aligned} \tag{13.3.3}$$

where $0 \leq \hat{\theta}_1 \leq 1, 0 \leq \hat{\tau}_1 \leq 1$, and $\hat{\theta}_2 = 1 - \hat{\theta}_1, \hat{\tau}_2 = 1 - \hat{\tau}_1$. Further, under the assumption of independence of A and B, $E(f_{ij}) = n\theta_i\tau_j$, and we can estimate $E(f_{ij})$ by $n\hat{\theta}_i\hat{\tau}_j = (f_{i\cdot} \times f_{\cdot j})/\ n$. Now we have to replace $\theta_1,\ \theta_2,\ \tau_1,\ \tau_2$ in (13.3.2) by their estimators given in (13.3.3), since $\theta_1,\ \theta_2,\ \tau_1,\ \tau_2$ are unknown. This then results in the statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\left[\frac{(f_{ij} - f_{i\cdot}f_{\cdot j}/n)^2}{f_{i\cdot}f_{\cdot j}/n}\right] \tag{13.3.4}$$

Algebraically, for this case of a *2 × 2 contingency table*, (13.3.4) can also be computed by

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{21}f_{12})^2}{f_{1\cdot}f_{2\cdot}f_{\cdot 1}f_{\cdot 2}} \tag{13.3.5}$$

For large n, χ^2 has approximately the chi-square distribution with $4 - 2 - 1 = 1$ degree of freedom. This is because the number of unknown parameters, c, to be estimated is two (one θ and one τ), and not four; that is, because once we have $\hat{\theta}_1$ and $\hat{\tau}_1$ we immediately have $\hat{\theta}_2 = 1 - \hat{\theta}_1$ and $\hat{\tau}_2 = 1 - \hat{\tau}_1$. Thus there are three constraints on sample f_{ij}'s, namely $\sum_i \sum_j f_{ij} = n,\ \hat{\theta}_1 = f_{1\cdot}/n = \sum_{j=1}^{2} f_{1j}/n$, and $\hat{\tau}_1 = f_{\cdot 1}/n = \sum_{i=1}^{2} f_{i1}/n$, accounting for the loss of three degrees of freedom; that is, the degrees of freedom for this case are $4 - 3 = 1$.

For this 2×2 case, the quantity χ^2, given by either (13.3.4) or (13.3.5), can then be used for testing the hypothesis that the A and B classifications are independent. That is, the sample frequencies displayed in Table 13.3.2 can be regarded as having probability distribution (13.3.1), where the probabilities of the four classes, $A \cap B$, $A \cap \bar{B}$, $\bar{A} \cap B$, and $\bar{A} \cap \bar{B}$, are as shown in Table 13.3.1 and $\theta_1, \theta_2, \tau_1, \tau_2$ are unknown. As the $\hat{\theta}_i$, and $\hat{\tau}_j$ are random variables, the statistic as now given by (13.3.4) or (13.3.5) has, for large n, the χ_1^2 distribution. Thus, if the observed value of χ^2, determined from (13.3.4) or (13.3.5) exceeds $\chi^2_{1,\alpha}$, we reject the hypothesis that the A and B classifications are independent at the $100\alpha\%$ level of significance and conclude that there is a significant degree of dependence between the A and B classifications at that same level of significance.

Example 13.3.2 (Nylon bar's brittleness and heat treatment) *In an experiment on bars of nylon, 800 randomly chosen bars were found to be such that 360 of them had been subjected to a* $60^\circ C$ *heat treatment and 440 to a* $90^\circ C$ *heat treatment. Each of the bars*

Table 13.3.4 Effect of heat treatment on the brittleness of nylon bars.

	Brittle B	Nonbrittle $\bar{B}$	
60°	(132)	(228)	
A	77	283	360
90°	(162)	(278)	
$\bar{A}$	217	223	440
	294	506	800

was then further classified as brittle or nonbrittle, with results shown in Table 13.3.4. The problem is to test the hypothesis that brittleness is independent of heat treatment.

Solution: The observed frequencies, $f_{11}, f_{12}, f_{21}, f_{22}$, for this experiment are the numbers 77, 283, 217, and 223, respectively. The estimate of $E(f_{11})$, for example is $n\hat{\theta}_1\hat{\tau}_1 = n(f_{1.}/n)(f_{.1}/n) = f_{1.}\, f_{.1}/n$. Thus, as shown in parentheses in Table 13.3.4, the expected frequencies of f_{11}, f_{12}, f_{21}, f_{22} are estimated to be 132, 228, 162, and 278, respectively. The observed value of χ^2 from these observations is

$$\chi^2 = \frac{(77-132)^2}{132} + \frac{(283-228)^2}{228} + \frac{(217-162)^2}{162} + \frac{(223-278)^2}{278} = 65.7$$

We could, of course, use the alternative formula (13.3.5) to find

$$\chi^2 = 800[(77)(223) - (217)(283)]^2/[(360)(440)(294)(506)] = 65.7$$

The value of $\chi^2_{1,.05}$, the critical value of chi-square with one degree of freedom at the 5% level of significance, is 3.84, which is less than the observed value 65.7. Therefore, at the 5% significance level, there is ample evidence to reject the null hypothesis of independence between heat treatment and brittleness, as suggested by the inspection of the data.

13.3.3 The $r \times s$ Contingency Table

The results in Sections 13.3.1 and 13.3.2 extend in a straightforward manner to the case of testing $r \times s$ contingency tables for independence. We have rs mutually exclusive and exhaustive classes, $A_i \cap B_j$, i =1, ..., r, j=1, ..., s. We assume that the A and B classifications are independent, so the probabilities of events $A_i \cap B_j$ are equal to $\theta_i\tau_j$, where $\theta_i = P(A_i)$ and $\tau_j = P(B_j)$ are all positive with $\theta_1 + \cdots + \theta_r = 1$, and $\tau_1 + \cdots + \tau_s = 1$. It is important to keep in mind that given any $(r-1)$ θ_i, the remaining θ is determined. For example, if we know the values of $(\theta_1, \ldots, \theta_{r-1})$, then $\theta_r = 1 - (\theta_1 + \cdots + \theta_{r-1})$. The situation is similar when dealing with τ_j. If n independent trials are performed, let f_{ij} be *the number of outcomes in class* $A_i \bigcap B_j$, where $\sum_{j=1}^{s} \sum_{i=1}^{r} f_{ij} = n$. Then the f_{ij} are random variables having the multinomial distribution

$$\frac{n!}{f_{11}!f_{12}!\ldots!f_{rs}!}(\theta_1\tau_1)^{f_{11}}(\theta_1\tau_2)^{f_{12}}\ldots(\theta_r\tau_s)^{f_{rs}} \qquad (13.3.6)$$

We then have the following:

Theorem 13.3.2 *If the θ_i and τ_j are known and if n is large, the quantity*

$$\chi^2 = \sum_{j=1}^{s} \sum_{i=1}^{r} \left[\frac{(f_{ij} - n\theta_i\tau_j)^2}{n\theta_i\tau_j} \right] \tag{13.3.7}$$

has approximately a chi-square distribution with rs −1 degrees of freedom.

However, if the θ_i and τ_j are unknown and are replaced by the estimators $f_{i\cdot}/n$ and $f_{\cdot j}/n$, respectively, where $f_{i\cdot} = f_{i1} + \cdots + f_{is}$ and $f_{\cdot j} = f_{1j} + \cdots + f_{rj}$, then we have

For large n,

$$\chi^2 = \sum_{j=1}^{s} \sum_{i=1}^{r} \left[\frac{(f_{ij} - f_{i\cdot}f_{\cdot j}/n)^2}{f_{i\cdot}f_{\cdot j}/n} \right] \tag{13.3.8}$$

has approximately a chi-square distribution with $rs - [(r-1) + (s-1)] - 1 = (r-1)(s-1)$ degrees of freedom. The result above holds since the number of constants in χ^2 to be estimated, as given by (13.3.7), are $c = (r-1) + (s-1)$, for $(r-1)$ θ_i's and $(s-1)$ τ_j's.

Example 13.3.3 (Effects of drug on learning ability) *A clinical psychologist wants to evaluate the effects of four different drugs (D_1, D_2, D_3, D_4) on the ability (high, average, low) to learn some unfamiliar material on a particular topic. One hundred subjects were selected randomly and classified according to the drug taken during the past six months and their ability to learn the unfamiliar material given to them. The data collected are shown in Table 13.3.5. The numbers in parentheses are the expected cell frequencies.*

Test, at the 5% level of significance, the null hypothesis that the types of drugs and learning ability are independent.

Table 13.3.5 Results of an experiment on effect of certain drugs on one's learning ability.

	Days				
Ability	D_1	D_2	D_3	D_4	Total
High	(5.04) 5	(5.04) 6	(6.09) 6	(4.83) 4	21
Average	(11.52) 7	(11.52) 11	(13.92) 14	(11.04) 16	48
Low	(7.44) 12	(7.44) 7	(8.99) 9	(7.13) 3	31
Total	24	24	29	23	100

Solution: Using (13.3.8), the observed value of χ^2 is given by

$$\chi^2 = \frac{(5-5.04)^2}{5.04} + \frac{(6-5.04)^2}{5.04} + \cdots + \frac{(3-7.13)^2}{7.13} = 9.429$$

In this example, the degrees of freedom for χ^2 are $(r-1)(s-1) = (3-1)(4-1) = 6$. The value of $\chi^2_{6;.05}$, the critical value of the chi-square with six degrees of freedom at the 5% level of significance, is 12.5916. Since this value is greater than the observed value of the χ^2-statistic, 9.429, we do not reject the null hypothesis of independence between the types of drug involved and learning ability.

Example 13.3.4 An educator wants to learn whether a*student's academic achievement depends upon his/her field (sciences or arts) of interest. A random sample of 180 students was selected and classified according to their grades in certain science courses and arts courses. The results obtained are summarized in Table 13.3.6. Test at the 5% level of significance the hypothesis that student grades are independent of their interest.*

Solution:

MINITAB

1. Enter the data in a MINITAB worksheet as shown below

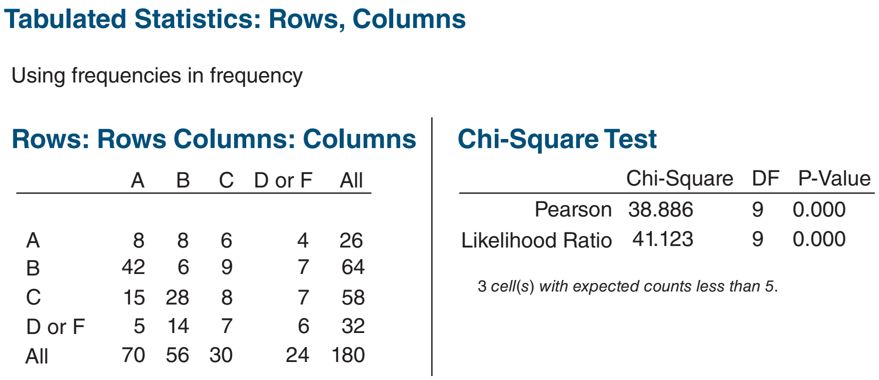
↓	C1-T	C2-T	C3
	Rows	Colums	Frequency
1	A	A	8
2	B	A	42
3	C	A	15
4	D or F	A	5
5	A	B	8
6	B	B	6
7	C	B	28
8	D or F	B	14
9	A	C	6
10	B	C	9
11	C	C	8
12	D or F	C	7
13	A	D or F	4
14	B	D or F	7
15	C	D or F	7
16	D or F	D or F	6

2. Select **Stat** > **Tables** > **Cross Tabulation and Chi-Square**.
3. In the dialog box, select **raw data (categorical variables)** from the pull down menu and type C1, C2, and C3 in boxes next to **Rows**, **Columns**, and **Frequencies**. Select any other desired options available in this dialog box. For example in this

Table 13.3.6 Data on students taking science and arts courses.

		Science course				
	Grades	A	B	C	D or F	Total
Arts course	A	8	8	6	4	26
	B	42	6	9	7	64
	C	15	28	8	7	58
	D or F	5	14	7	6	32
	Total	70	56	30	24	180

problem, we select chi-square and check appropriate items in the new dialog box. Then click **OK**. The MINITAB output that appears in the Session window is shown below.



```
Tabulated Statistics: Rows, Columns

Using frequencies in frequency

Rows: Rows Columns: Columns           Chi-Square Test
           A   B   C  D or F  All                   Chi-Square  DF  P-Value
                                              Pearson  38.886     9  0.000
A          8   8   6     4    26     Likelihood Ratio  41.123     9  0.000
B         42   6   9     7    64
C         15  28   8     7    58     3 cell(s) with expected counts less than 5.
D or F     5  14   7     6    32
All       70  56  30    24   180
```

Since the p-value is 0.00, which is less than the level of significance 0.05, we reject the null hypothesis that student grades are independent of their field of interest.

USING R

The following manual R code can be used to conduct the chi-squared goodness of fit for a 4×4 contingency table.

```
#Assign data
Rows = c(‘A’, ‘B’, ‘C’, ‘D or F’, ‘A’, ‘B’, ‘C’, ‘D or F’, ‘A’, ‘B’, ‘C’, ‘D or F’,
         ‘A’, ‘B’, ‘C’, ‘D or F’)
Columns = c(‘A’, ‘A’, ‘A’, ‘A’, ‘B’, ‘B’, ‘B’, ‘B’, ‘C’, ‘C’, ‘C’, ‘C’, ‘D or F’,
            ‘D or F’, ‘D or F’, ‘D or F’)
Frequency = c(8,42,15,5,8,6,28,14,6,9,8,7,4,7,7,6)

#Make a data.frame
Table = data.frame(Rows, Columns, Frequency)

#Create a 2×2 table similar to Table 13.3.6
Tab.data = xtabs(Frequency ~ Rows+Columns, data=Table)

#chi-squared test for independence
chisq.test(Tab.data) #R output

Pearson's Chi-squared test
data: Tab.data
X-squared = 38.886, df = 9, p-value = 1.208e-05
```

PRACTICE PROBLEMS FOR SECTION 13.3

1. A random sample of 106 engineers was asked their starting salaries, and whether they graduated from a public or a private school. The results are given below:

School	\$40,000–\$49,999	\$50,000–\$59,999	\$60,000–\$69,999	Over \$70,000	Total
Public	12	18	8	6	44
Private	8	24	14	16	62
Total	20	42	22	22	106

   Test at the 5% level of significance the null hypothesis that the starting salaries of engineers are independent of the type of school.

2. Recently, a new strain of virus that causes influenza was detected. A new vaccine to counter this strain was developed and tested using two selected groups. One group was inoculated with the new vaccine, while the other group did not receive any vaccination. The results showing how many people became sick and did not become sick from the two groups are given below:

	Vaccinated	Not vaccinated	Total
Got sick	45	95	140
Did not get sick	125	65	190
Total	170	160	330

   Test at the 1% level of significance that the two factors “vaccination” and “sickness” are independent.

3. A social worker was interested in finding the effect of education among women on their married life. The data given below are the levels of their education and the number of years they remained married to their first husband. Test at the 5% level of significance the null hypothesis that the factors “education” and “marriage” are independent.

	<10	$10–< 20$	$20–<30$	≥ 30	Total
High school	37	31	20	12	100
Undergraduate degree	25	40	15	20	100
Graduate degree	20	25	35	20	100
Professional degree	10	8	30	52	100
Total	92	104	100	104	400

4. A medical team conducted an experiment to observe if hypertension is dependent on drinking habits. The data below give information on 250 individuals. Test the hypothesis at the 5% level of significance that having or not having hypertension is independent of drinking habits.

Hypertension	Nondrinkers	Moderate drinkers	Heavy drinkers
Yes	20	44	42
No	80	36	28

5. A social worker conducted an experiment to study if the level of education among immigrants to the United States is dependent on the region of migration. The following data give the level of education of 815 immigrants. Test the hypothesis at the 1% level of significance that the level of education among immigrants is independent of their region of migration.

Level of education	Asian countries	European countries	Middle Eastern ountries	African countries
Highly educated	90	80	40	30
Well educated	120	120	50	35
Not well educated	70	50	60	70

6. The following data give the income level for 400 married couples and the number of children that they have. Test the hypothesis at the 5% level of significance that the number of children is independent of the level of income.

Number of children	0–2	3	4	5 or more
Less than $75K	32	30	16	10
$75K-$150K	38	44	10	8
Over $150K	130	66	9	7

13.4 CHI-SQUARE TEST FOR HOMOGENEITY

In Section 13.2, we discussed the testing of the goodness-of-fit hypothesis. That is, if a given sample is classified into k categories, then we verify that the sample comes from some theoretical population, which may or may not be completely specified. In particular, we discussed the hypothesis of fitting a multinomial model to the given data. In Section 13.3, we extended this to the $r \times s$ contingency table case ($k = rs$), that is when one random sample of size n is taken from *one population* and each element of the sample is classified with respect to two criteria of interest. In this section, we discuss a problem that may be considered as an extension of the problem discussed in Section 13.3. That is, for the problem of this section, random samples from s populations are taken and elements of each sample are classified into r categories of interest.

Suppose that we have s populations, and we wish to test a hypothesis that *the s populations are homogeneous* with respect to some characteristic of interest. By “homogeneous” we mean that each of the s populations has the same distribution with respect to some characteristic of interest. To formalize this, suppose we have s populations, say $B_1, \ldots, B_s$, and that for each population we are interested in the proportions of the population that belongs to one of the levels $A_1, \ldots, A_r$ of the characteristic of interest, A. That is, for each B_j we are interested in

$$P(A_i|B_j), \quad i = 1, \ldots, r$$

Then, the s populations $B_1, \ldots, B_s$ are homogeneous if

$$P(A_i|B_j) = P(A_i), \qquad i = 1, \ldots, r \tag{13.4.1}$$

for each $j = 1, \ldots, s$.

Equation (13.4.1) implies that the s populations behave in the same fashion with respect to the characteristic of interest A. We now notice that (13.4.1) says that A and B are independent. To test independence of A and B (i.e., the B populations are homogeneous with respect to the characteristic of interest A), we take random samples from each of the populations B_j, $j = 1, \ldots, s$. We next classify the members of the samples from B_j as to which of the levels A_i of A they belong, $i = 1, \ldots, r$, obtaining an $r \times s$ table of f_{ij}'s, say, where f_{ij} is the number of members of the jth sample obtained from the population B_j that are classified as belonging to level A_i. This is similar in format to the $r \times s$ contingency table. It is not surprising, then, that the test for independence of A and B is equivalent to testing the homogeneity of populations $B_1, \ldots, B_s$. Hence, we may use the chi-square goodness-of-fit statistic of Section 13.3, namely

$$\chi^2 = \sum_{j=1}^{s} \sum_{i=1}^{r} \left[\frac{(f_{ij} - f_{i\cdot} f_{\cdot j}/n)^2}{f_{i\cdot} f_{\cdot j}/n} \right] \tag{13.4.2}$$

We illustrate this procedure with the following example.

Example 13.4.1 (Comparing economic conditions in New England states) *Suppose we have obtained six random samples, one from each of the New England states Connecticut (CT), Maine (ME), Massachusetts (MA), New Hampshire (NH), Rhode Island (RI), and Vermont (VT). Suppose that each sample is separately classified into four mutually exclusive categories according to income level and that the data obtained is as given in Table 13.4.1 (numbers in parentheses are expected frequencies). Test at the 5% level of significance that all the six states are economically homogeneous.*

Table 13.4.1 Data on income levels in states of New England.

	CT	ME	MA	NH	RI	VT	Total
Under \$30K	(10.62)	(6.56)	(13.38)	(6.03)	(5.64)	(5.77)	
	6	10	8	6	6	12	48
\$30K-under \$50K	(21.91)	(13.52)	(27.59)	(12.44)	(11.63)	(11.90)	
	20	18	22	11	12	16	99
\$50K-under \$100K	(29.88)	(18.44)	(37.62)	(16.97)	(15.86)	(16.23)	
	35	17	42	18	13	10	135
Over \$100K	(18.59)	(11.47)	(23.41)	(10.56)	(9.87)	(10.10)	
	20	5	30	11	12	6	84
Total	81	50	102	46	43	44	366

Solution: We wish to test at the 5% level of significance, the hypothesis

H_0: The six New England states are homogeneous with respect to income level
H_1: The six New England states are not homogeneous with respect to income level

Now the discussion in this section has shown this problem to be equivalent to testing the following hypothesis:

H_0: The factors "income level" and "state" are independent
H_1: The factors "income level" and "state" are not independent

Hence, we use the goodness-of-fit statistic defined by

$$\chi^2 = \sum_{j=1}^{6} \sum_{i=1}^{4} \left[\frac{(f_{ij} - f_{i\cdot} f_{\cdot j}/n)^2}{f_{i\cdot} f_{\cdot j}/n} \right] \qquad (13.4.3)$$

Here $n = 366$ and the f_{ij} are given in Table 13.4.1. Hence, we would reject at the 5% level of significance the null hypothesis that the six New England states are homogeneous with respect to income level if the observed χ^2 is such that

$$\chi^2 > \chi^2_{15;0.05} = 24.996$$

since $(r-1)(s-1) = (4-1)(6-1) = 15$.

The observed value of χ^2 statistic is, for the data of this example,

$$\chi^2 = \frac{(6-10.62)^2}{10.62} + \cdots + \frac{(6-10.10)^2}{10.10} = 29.31832 > 24.996$$

so we reject the null hypothesis of homogeneity of the six New England states with respect to income level.

PRACTICE PROBLEMS FOR SECTION 13.4

1. An outpatient clinic at a city hospital conducted an experiment to treat a viral infection using three drugs: Acetaminophen, Motrin, and Ibuprofen. Each drug was administered on 80 patients. The results are shown below. Test at the 1% level of significance the null hypothesis that the three drugs are equally effective.

Level of relief	Acetaminophen	Ibuprofen	Motrin
Little or no relief	18	22	12
Some relief	26	18	30
Complete relief	36	40	38
Total	80	80	80

2. A company operates five machines in three shifts daily. The five machines are made by five different manufacturers, and the maintenance manager of the company keeps a log of breakdowns over the past three months. The results are shown below. Test at the 5% level of significance the null hypothesis that the five machines are equally prone to breakdowns.

Shift	Machine 1	Machine 2	Machine 3	Machine 4	Machine 5	Total
Morning	8	9	6	10	11	44
Evening	9	12	14	12	15	62
Night	7	8	13	10	16	54
Total	24	29	33	32	42	160

3. Three hundred CEOs from five different industries were asked whether the higher interest rates will affect their hiring during the next two fiscal years. The results obtained are shown below. Test at the 10% level of significance the null hypothesis that all five industries are equally affected by a higher interest rate.

	Food	Manufacturing	Pharmaceutical	Insurance	General	Total
Yes	15	35	25	15	8	98
No	35	25	45	45	52	202
Total	50	60	70	60	60	300

4. An insurance company wished to study the areas of interest of medical graduates from medical schools, osteopathic medicine schools, and foreign medical schools in different specialties. The following data give information about 200 recent medical graduates. Test at the 5% level of significance the hypothesis that the three groups of medical students are homogeneous with respect to interest in different specialties.

Specialties	Family med	Internal med	Cardiology	Radiology	Other specialties
Medical school	5	10	15	18	10
Osteopathic school	25	12	10	12	6
Foreign med. school	35	24	7	5	6

5. The following data give information about income level and the marital status of 400 attendees of a concert organized by United Way for the benefit of Iraq War veterans. Test the hypothesis at the 5% level of significance that the three populations are homogeneous with respect to the marital status.

Income Level	Marital Status			
	Single	Married	Separated	Divorced
Less than \$100K	22	20	16	7
\$100K-\$250K	43	44	12	11
Over \$250K	130	66	20	9

6. The following data give the results of a study that was conducted to evaluate the effect of social status on the level of education. Test at the 5% level of significance if these data provide sufficient evidence that the three classes are homogeneous with respect to level of education.

Social status	Education level					
	MD/DO	Law	PhD	MS/MBA	Undergraduate	High school
Upper class	20	15	10	25	12	8
Middle class	18	25	8	20	14	12
Lower middle class	8	10	8	18	24	32

13.5 COMMENTS ON THE DISTRIBUTION OF THE LACK-OF-FIT STATISTICS

In the previous sections of this chapter, we have used some distribution results without full justification. We now proceed to discuss this aspect within the scope of this book.

We begin with a binomial random variable X, which is the number of trials in a sample of n independent trials that have the characteristic A. We know, by observing X, that $Y = n - X$ is the number of trials in this sample that has the characteristic $\bar{A}$.

Suppose that the probability, $P(A)$, of obtaining A in a single trial is θ so that

$$P(A) = \theta, \quad P(\bar{A}) = 1 - \theta \tag{13.5.1}$$

Recall that $E(X) = n\theta$ and $\mathrm{Var(X)} = n\theta(1-\theta)$. Then, using the Central Limit Theorem, the distribution of X when n is large can be approximated by the normal distribution with

$$\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \sim Z \tag{13.5.2}$$

where Z is a $N(0,1)$ random variable.

From Chapter 7, we know that the square (Z^2 in this case) of an $N(0,1)$ random variable is distributed as a χ_1^2 variable. Hence we may rewrite (13.5.2) in the form

$$\frac{(X-n\theta)^2}{n\theta(1-\theta)} \sim \chi_1^2 \quad \text{for large } n \tag{13.5.3}$$

Upon using the facts $1/\theta + 1/(1-\theta) = 1/\theta(1-\theta)$ and $Y = n - X$, with $X - n\theta = -[Y - n(1-\theta)]$, it is easily verified that the left-hand side of (13.5.3) may take the form

$$\frac{(X-n\theta)^2}{n\theta} + \frac{[Y-n(1-\theta)]^2}{n(1-\theta)} \tag{13.5.4}$$

Now let

$$X = f_1, \quad Y = f_2, \quad A = A_1, \quad \bar{A} = A_2, \quad \theta_1 = P(A) = P(A_1)$$

and $\theta_2 = P(\bar{A}) = P(A_2)$. Of course, f_i is the number of trials in the sample of n trials that result in A_i for $i = 1$ and 2. Since each f_i has the binomial distribution, we know that

$$E(f_i) = n\theta_i \tag{13.5.5}$$

We can put (13.5.3) and (13.5.4) together to obtain, for large n,

$$\sum_{i=1}^{2} \frac{(f_i - n\theta_i)^2}{n\theta_i} \cong \chi_1^2 \tag{13.5.6}$$

Note that $k = 2$, but since $f_1 + f_2 = n$, we then have $f_2 = n - f_1$, so that and if f_1 is observed, we automatically know f_2. Hence, the degrees of freedom for the approximate chi-square distribution is 1.

In the case of k characteristics $A_1, A_2, \ldots, A_k$, the respective frequencies, $f_1, f_2, \ldots, f_k$, have a multinomial distribution. We note that

$$\sum_{i=1}^{k} f_i = n \quad \text{or} \quad f_k = n - (f_1 + \cdots + f_{k-1}) \tag{13.5.7}$$

Indeed, the probability function of $f_1, \ldots, f_{k-1}$ is

$$p(f_1, \ldots, f_{k-1}) = \frac{n!}{\prod_{i=1}^{k-1} f_i! \left[n - \sum_{i=1}^{k-1} f_i\right]!} {\theta_1}^{f_1} \ldots {\theta_{k-1}}^{f_{k-1}} \left(1 - \sum_{i=1}^{k-1} \theta_i\right)^{\left(n - \sum\limits_{i=1}^{k-1} f_i\right)} \tag{13.5.8}$$

for $0 \le f_i \le n$ and $0 \le \sum_{i=1}^{k-1} f_i \le n$.

Thus, just as in the case of $k = 2$ ($k - 1 = 1$), where we are dealing with a binomial random variable, it can be proved that the distribution (13.5.8) is well approximated *for* *large n* by a certain $(k-1)$-dimensional normal distribution (see Chapter 6 for a discussion of the two-dimensional normal distribution, called the bivariate normal). It can be proved that, for large n, we have

$$\sum_{i=1}^{k} \frac{(f_i - n\theta_i)^2}{n\theta_i} \sim \chi^2_{k-1} \tag{13.5.9}$$

where $f_k = n - \sum_{i=1}^{k-1} f_i$, $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$.

This result, for $k = 2$, was presented at the beginning of this section.

Now, when $n\theta_i = E(f_i)$, $i = 1, \ldots, k$, are unknown and we make c estimates, $c < k-1$ based on the sample to furnish a set of estimates of the $n\theta_i$'s, we are placing c further restrictions on the f_i's (e.g. see Section 13.3.2 or 13.3.3). These restrictions cause the loss of c additional degrees of freedom, in addition to the loss of the one degree of freedom imposed by the restriction $\sum_{i=1}^{k} f_i = n$.

13.6 CASE STUDIES

Case Study 1 (*Data on vitamin D and PTH levels during pregnancy*)
In this study the authors—James E. Haddow, Glen E. Palomaki, Geralyn Lambert-Messerlian, Louis M. Neveux, Jacob A. Canick, David M. Grenache, and Jun Lu—examine the relationship between 25(OH)D and the parathyroid hormone (PTH) concentration in first-trimester pregnant women from New England and their overall vitamin D levels in comparison to earlier reports.

The authors retrieved residual sera stored at −20°C after routine first-trimester Down syndrome screenings in 2008 of 432 African American and 587 Caucasian women, distributed evenly over 12 months. Samples were tested for 25(OH)D and PTH. The collected data (Table 13.6.1 Case Study 13.1) is available on the book website: www.wiley.com/college/gupta/statistics2e. These data provide the vitamin D and PTH levels during pregnancy. (Legend: Race: 1-black, 2-Caucasian; Season: 1-spring, 2-summer, 3-fall, 4-winter; Cigarettes-number smoked per day) (Source: Data used with permission).

Select a random sample of size 100 from the data in Table 13.6.1.

(a) Construct a contingency table for race and PTH (pmol/L) concentration levels. Use four categories of PTH concentration: category 1 (0-under 9), category 2 (9-under 12.5), category 3 (12.5-under 16), category 4 (16 or more). Test at the 5% level of significance the hypothesis that race and PTH (pmol/L) concentration levels are independent.

Table 13.6.1 Data comparing outcomes of matches 1 and 2 with those of matches 4 and 5.

	Matches 1 and 2	Matches 4 and 5	Total
Home wins	120	33	153
Away wins	68	21	89
Total	188	54	242

(b) Construct a contingency table for race and vitamin D (nmol/L) levels. Use three categories of vitamin D (nmol/L) levels: category 1 (0-under 35), category 2 (35-under 70), category 3 (70 or more). Test at the 1% level of significance the hypothesis that race and vitamin D (nmol/L) levels are independent.
(c) Construct a contingency table for seasons and vitamin D (nmol/L) levels. Use three categories of vitamin D (nmol/L) levels: category 1 (0-under 35), category 2 (35-under 70), category 3 (70 or more). Test at the 1% level of significance the hypothesis that seasons and vitamin D (nmol/L) levels are independent.
(d) Construct a contingency table for race and weight in lb. Use three categories of weight: category 1 (0-under 135), category 2 (135-under 160), and category 3 (160 or more). Test at the 1% level of significance the hypothesis that race and weights are independent.

Case Study 2 (*Home advantage in sport competitions*[1])
The Davis Cup is the annual tennis tournament for men's international teams. The International Tennis Federation runs cup tournaments during which teams of players compete in a knockout format. The world's 16 best national teams are selected and compete for the Davis Cup. The rounds consist of two singles matches, followed by a double match and, if necessary, two more singles matches. The country team that wins three matches is the winner. Tables 13.6.1 and 13.6.2 show the scores for matches from 1900 to 2006. Each match was recorded as a home or away win. Two separate data group were collected: data comparing outcomes of matches 1 and 2 with those of matches 4 and 5; data comparing outcomes of matches 1 and 2 with that of match 5 (Source: Data are used with permission). Use the Chi-square test to analyze the data in Tables 13.6.1 and 13.6.2, and state your conclusions about home advantage or disadvantage in the Davis Cup tournaments.

Table 13.6.2 Data comparing outcomes of matches 1 and 2 with that of match 5.

	Matches 1 and 2	Match 5	Total
Home wins	120	8	128
Away wins	68	13	81
Total	188	21	209

Case Study 3 (*Data on vitamin D and PTH levels*, continued) Select a random sample of size 150 from the data (Table 13.6.1) available on the book website: www.wiley.com/college/gupta/statistics2e.

Develop various contingency tables to test the two populations: African-American women and Caucasian women are homogeneous with respect to some other variables. Explain how this case study differs from case study 1.

13.7 USING JMP

This section is not included in the book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

[1] Source: Gayton et al. (2009).

Review Practice Problems

1. Suppose that a coin is tossed 1000 times with the result that 462 heads and 538 tails are obtained. Are these results consistent with the hypothesis that the coin is true at the 5% level of significance? (Use the chi-square test.)

2. Five thumbtacks of a certain type were thrown 200 times and the number of tacks of the five falling point up in each throw was counted. The experimental results are shown below.

Tacks falling point up, X	Frequency
0	5
1	27
2	41
3	67
4	43
5	17
Total	200

   (a) Estimate the probability θ that a tack falls point up.
   (b) Test the null hypothesis that the distribution of X, the number of tacks falling point up, has the binomial distribution $\binom{5}{x}\theta^x(1-\theta)^{5-x}$, $x = x = 0, 1, \ldots, 5$, using the estimate of θ found in (a). State α and the observed level of significance.

3. Using the method of Example 13.2.4, fit a normal distribution to the data of Problem 14 of Review Practice Problems in Chapter 2 and test the null hypothesis at the 5% level of significance that the data behave as a sample from a normal population. Find the observed level of significance.

4. Houses with and without air conditioners on nine different streets in a certain city are shown below (from Brownlee, 1957). Test (at the 5% level) the null hypothesis that having an air conditioner is independent of street locations of the houses.

Street	With A/C	Without A/C
1	5	18
2	8	35
3	18	25
4	3	38
5	17	24
6	11	31
7	25	17
8	19	20
9	18	18

5. A small roulette wheel was spun 380 times and yielded frequencies shown below for the 38 roulette numbers (taken in pairs). Test the hypothesis that the wheel is true, using the chi-square test at the 1% level of significance.

Spins	Frequency	Spins	Frequency	Spins	Frequency
0–00	24	13–14	29	25–26	21
1–2	16	15–16	21	27–28	14
3–4	19	17–18	17	29–30	25
5–6	19	19–20	25	31–32	23
7–8	25	21–22	18	33–34	16
9–10	10	23–24	20	35–36	16
11–12	22				
Total					380

6. Pieces of vulcanite were examined according to porosity and dimensional defects, and the results are shown below (data from Hald, 1952). Test the hypothesis that the two criteria of classification are independent. Specify α.

	Porous	Nonporous
With defective dimensions	142	331
Without defective dimensions	1233	5099

7. In field tests of mine fuses, 216 of each of the two types of fuses A and B, chosen at random from large lots, were buried, and then simulated tanks ran over them. The number of "hits" and "not hits" was recorded for each type of fuse, with the results shown below (from Ordnance Corps Pamphlet ORD P 20 = 111). Are the proportions of hits for the two types of fuses significantly different at the 5% level of significance?

Fuse type	Hit	Not Hit	Total
A	181	35	216
B	160	56	216
Total	341	91	432

8. The lateral deflection and range in yards obtained in firing 75 rockets are shown below (from Crow et al. (1955)]). Test at the 5% level of significance the hypothesis that lateral deflection and range are independent.

	Lateral deflection (yards)			
Range(yards)	−250 to −51	−50 to +49	50 to 199	Total
0–1199	5	9	7	21
1200–1799	7	3	9	19
1800–2699	8	21	6	35
Total	20	33	22	75

9. A study was performed on the effect of time of work on quality of work in a certain plant. It was the practice in the plant for a crew to change shifts once a month. A study of three months of operations by one crew that remained intact for the entire period showed the numbers of defective and nondefective items produced. The data are given below (from Duncan, 1958). Do you conclude from these data that the time of work significantly affects the quality of the works? Use the 0.05 level of significance. Justify your answer.

Shift	Defective	Nondefective
1 (8:00–4:00)	52	921
2 (4:00–12:00)	61	902
3 (12:00–8:00)	73	851

10. The number of contaminated tablets were counted for 720 samples with each sample consisting of 100 tablets. The results are shown below. Fit a Poisson distribution to this data and test for goodness of fit. Specify α.

Contaminated tablets (X_i) in a sample	0	1	2	3	4	5	6	7	8	9	10
Observed sample frequency (f_i)	116	194	184	115	63	24	12	6	3	2	1

11. A certain strain of guinea pig is such that 75% of its progeny are born with white eyes and 75% of its progeny are born with webbed feet. A sample of 160 piglets from newly born litters is classified as shown below. Test whether the modes of classification are independent. Use $\alpha = 0.05$.

	Webbed feet	Non-webbed feet
White eyes	94	33
Colored eyes	28	5

12. The data given below show the classification of 357 persons by race who visit a physician's office a certain number of times over a period of one year. The results are shown below. Test, at the 5% level of significance, the hypothesis that the four races are homogeneous with respect to the number of visits to a physician's office. Find the observed level of significance.

Number of visits	White	African American	Hispanic	Asian
0–1	50	34	38	40
2–4	28	35	42	30
5 or more	12	18	16	14

13. A random sample of bulbs is taken from lots manufactured in the United States, Canada, and Mexico. Then, from each sample, the defectives and nondefective bulbs are separated. The results are shown below. Test at the 1% level of significance the hypothesis that the quality of bulbs in all three lots is the same. Find the observed level of significance.

	United States	Canada	Mexico
Nondefective	320	280	295
Defective	15	14	17

14. The direct investment by US companies in millions of dollars in five different countries (A, B, C, D, and E) during 1995, 2000, and 2005 are as shown below. Do these data provide sufficient evidence to indicate that the US investments over time in these countries differ significantly? Use $\alpha = 0.05$. Find the observed level of significance.

	1995	2000	2005
A	551	1110	1143
B	288	637	648
C	598	1888	1924
D	387	763	417
E	519	737	942

15. The following data set gives the scores of 50 students of an engineering exam. The results are shown below. Test at the 1% level of significance that these data follow a normal distribution.

79	81	82	74	86	92	95	87	70	78
79	88	89	85	87	92	96	91	83	83
77	76	84	87	88	91	90	92	94	96
94	91	90	77	81	76	78	83	85	87
80	87	86	85	84	86	88	86	87	85

16. The data below give the frequency distribution of ages for 100 students selected randomly from a large university. Test at the 5% level of significance that the data in the table follow a multinomial distribution, with $\theta_1 = 0.35,\ \theta_2 = 0.25, \theta_3 = 0.15,\ \theta_4 = 0.10, \theta_5 = 0.10,\ \theta_6 = 0.05$.

Class	1	2	3	4	5	6
Class enrollment	18–23	24–29	30–35	36–41	42–47	48 & over
Frequency	36	30	12	8	6	8

17. The data given below provide the frequency distribution of daily traffic violation tickets issued by the city police over a period of eight weeks. Test at the 5% level of significance the hypothesis that the number of traffic tickets is the same for each of the past eight weeks.

Week	1	2	3	4	5	6	7	8
Violation tickets	19	25	21	24	15	28	20	15

18. The data below provide the number of customers entering in a bank during morning hours (9 am–12 noon) on each day of a given week. Do these data provide sufficient evidence that the number of customers entering in the bank is not the same on the different working days of the week? Use $\alpha = 0.05$.

Weekday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Customers	60	72	90	75	95	92

19. The data below gives the time (in minutes) elapsed between the admission of patients to a certain hospital. Can we conclude, at the 5% level of significance, that these data follow an exponential distribution?

15	18	78	5	16	58	78	60	40	60
55	49	28	30	39	18	10	12	17	21
24	23	27	14	11	8	28	36	69	53
59	72	66	49	46	42	37	36	47	45
50	59	55	64	67	23	25	17	15	18

20. An experiment of tossing a certain coin four times is repeated 100 times and the number of heads appearing in each experiment is recorded and shown below. Test at the 5% level of significance the hypothesis that the coin used is unbiased.

Number of heads	0	1	2	3	4
Frequency	8	18	35	25	14

21. A computer assembling company buys all its memory cards from three manufacturers M_1, M_2, M_3. The quality control manager of the company decided to learn more about the quality of the memory cards purchased from the three manufacturers, and thus took a random sample of 500 cards from the shipments received from each manufacturer and classified them according to whether the card is defective or non-defective. The results are shown below. Test, at the 5% level of significance, the hypothesis that the quality of the memory cards is independent of the manufacturer.

	M_1	M_2	M_3
Defective	32	52	28
Nondefective	468	448	472

22. To determine whether there is any relationship between the school attended to earn an engineering degree and success at a job, a sample of 300 engineers who earned their degree from different schools S_1, S_2, S_3 is randomly selected, and then each engineer is classified according to his/her success (high, average, low) and the school from where they earned their degree. The data are shown below. Test, at the 1% level of significance, the hypothesis that the success at a job is independent of the school where one earns his/her degree.

	S_1	S_2	S_3
High	25	39	23
Average	42	67	32
Low	21	33	18

23. A pharmaceutical company wants to study the form of a drug that is possibly more effective. Three different forms, namely tablet, suspension, and injection were prescribed randomly to 190 patients. After using that drug for four weeks, its effectiveness is observed and the data obtained are shown below. Test, at the 1% level of significance, the hypothesis that the effectiveness of the drug is independent of the form of the drug.

Effectiveness	Tablet	Suspension	Injection
High	20	23	28
Average	32	20	22
Low	18	15	12

24. A random sample of 500 teenagers is selected and classified according to age and the number of accidents he/she has had over a given period of time. The data are shown below. Test, at the 1% level of significance, the hypothesis that the age and number of accidents are independent.

	Accidents			
Age	0	1	2	3 or more
16	50	20	10	5
17	84	49	20	7
18	40	48	32	10
19	48	62	10	5

Chapter 14

NONPARAMETRIC TESTS

The focus of this chapter is the development of some commonly used nonparametric procedures.

Topics Covered

- The one-sample and two-sample sign test
- The Mann–Whitney (Wilcoxon) W test for two samples
- Run tests: runs above and below the median and the Wald–Wolfowitz run test
- Spearman rank correlation

Learning Outcomes

After studying this chapter, the reader will be able to

- Use a nonparametric method for testing hypotheses about a location parameter when the sample is drawn from one population.
- Use a nonparametric method for testing hypotheses about location parameters when samples are drawn from two populations.
- Perform a nonparametric test of whether or not the sample at hand is a random sample.
- Investigate whether or not there is some association between two variables.

14.1 INTRODUCTION

In Chapter 9, we discussed various statistical tests based on the assumption that the samples involved are drawn from normal populations. There are many situations where we know little about the shape of the population distribution from which the samples are drawn, and in such cases the assumption of normality may be hazardous. There is a class

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

of statistical tests that are valid for samples from continuous population distributions of any shape. These are called *nonparametric tests*, and they are based on order statistics. In this chapter, we consider several of the simplest and most widely used of these tests. Throughout this chapter, we assume that the random variable of interest is a *continuous* random variable and the scale of measurement is at least *ordinal.*

14.2 THE SIGN TEST

14.2.1 One-Sample Test

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with an unknown median M and that we are interested in testing a hypothesis that the median takes some specified value. That is, the null and alternative hypotheses are

$$H_0\colon\ M = m_0 \quad \text{versus} \quad H_1\colon\ M < m_0 \tag{14.2.1a}$$

$$H_0\colon\ M = m_0 \quad \text{versus} \quad H_1\colon\ M > m_0 \tag{14.2.1b}$$

$$H_0\colon\ M = m_0 \quad \text{versus} \quad H_1\colon\ M \neq m_0 \tag{14.2.1c}$$

As in Chapter 9, the hypotheses (14.2.1a) and (14.2.1b) lead to one-tail tests and Equation (14.2.1c) leads to two-tail tests. The probability of the type I error, or the level of significance, is denoted by α, which takes some assigned value. In practice, α is usually assigned the values 0.1, 0.05, or 0.01. Recalling the definition of a population median, then under H_0 the null hypothesis, each $X_i, i = 1, \ldots, n$, has the same probability 1/2 of being greater than or less than m_0. In other words, $X_i - m_0$ has the same probability of being either positive (+) or negative (−). (This is why the test procedures of this section are called sign tests.) To test any one of the hypotheses, we define

$$Y_i = \begin{cases} 1, & \text{if } X_i - m_0 > 0 \\ 0, & \text{if } X_i - m_0 < 0 \end{cases}$$

for $i = 1, \ldots, n$. If any $X_i - m_0 = 0$, then we discard the corresponding observation and adjust the sample size n by reducing it by the number of observations discarded. Then, for the sample size n used, the test statistic is

$$R = \sum_{i=1}^{n} Y_i \tag{14.2.2}$$

and the observed value of the test statistic, say r, is the number of 1's among the Y_i. We note that under H_0 of Equation (14.2.1a), Y_i is Bernoulli, with $p_i = P(Y_i = 1) = p = 1/2$ so under H_0, $R = \sum_{i=1}^{n} Y_i$ is distributed as binomial with sample size n and $p = 1/2$, so that

$$P(R = r|H_0) = \binom{n}{r}\left(\frac{1}{2}\right)^n$$

Based on this result, the decision rules for the problem of Equations (14.2.1a) and (14.2.1c), with level of significance α are as follows:

(a) The hypotheses (14.2.1a) defines a left-sided hypothesis testing problem. Hence, reject H_0 in favor of H_1 if the observed value of R in Equation (14.2.2) is too small so that the probability of getting r or fewer 1s in a random sample of size n is less than or equal to α. In other words, reject H_0 if the p-value is less than or equal to α.

(b) The hypotheses (14.2.1b) defines a right-sided hypothesis testing problem. Hence, we reject H_0 in favor of H_1 if the observed value of R in Equation (14.2.2) is too large so that the probability of getting r or more 1s in a random sample of size n is less than or equal to α. In other words, reject H_0 if the p-value is less than or equal to α.

(c) The hypotheses (14.2.1c) defines a two-sided test so that we reject H_0 in favor of H_1 if the observed value of R, say r, is either significantly low or significantly high at the α level of significance. To formalize this, let $r' = \min(r, n-r)$, where r is the observed value of R. Obviously $r' \leq n/2$. Then, we reject H_0 if

$$P(r' < R < n - r') < 1 - \alpha$$

so that the p-value would then be (H_0 versus H_1 in Equation (14.2.1c))

$$\begin{aligned} p\text{-value} &= 1 - P\left(r' < R < n - r' | R \sim B\left(n, \tfrac{1}{2}\right)\right) \\ &= P(R \leq r'|H_0) + P(R \geq n - r'|H_0) = 2P(R \leq r'|H_0) \end{aligned} \tag{14.2.3}$$

Now $R \sim B(n, 1/2)$ under H_0, and the probability function of the $B(n,1/2)$ distribution is symmetric, that is,

$$P\left(x \middle| n, \frac{1}{2}\right) = P\left(n - x \middle| n, \frac{1}{2}\right)$$

Hence, reject H_0 if the p-value (14.2.3) is less than or equal to α, or if

$$\frac{1}{2}(p\text{-value}) = P(R \leq r'|H_0) = \sum_{r=0}^{r'} \binom{n}{r} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2} \tag{14.2.4}$$

We illustrate this method with the following example:

Example 14.2.1 (Job completion time) *A manager of a manufacturing company is interested in finding the median time taken by company technicians to complete a job. The manager takes a sample of 10 technicians and observes the time (in minutes) taken to complete the job, as follows:*

42 40 46 44 43 48 41 42 42 44

Test at the 5% level of significance that the median time taken by the company technician to complete a job is 45 minutes.

Solution: In this example, we wish to test the hypothesis

$$H_0\colon\ M = 45 \quad \text{versus} \quad H_1\colon\ M \neq 45$$

To test this hypothesis, we subtract 45 (the value under the null hypothesis) from each observation and assign the values 1 or 0 based on whether or not the difference is greater than 0 or less than 0. For the sample above, we start by recording y_i values, namely

$$0, 0, 1, 0, 0, 1, 0, 0, 0, 0$$

so that the observed value r of the test statistic is 2, the number of 1s.

Here $n = 10$, so that $r' = \min(r, n - r) = \min(2, 8) = 2$. Now using binomial tables, we have

$$p\text{-value} = 2 \times P(R \leq 2|n = 10, p = 0.5) = 2 \times 0.055 = 0.11$$

so that the p-value for this test is $0.11 > \alpha = 0.05$. We do not reject the null hypothesis. Recall that, for large sample size n, we could use the normal approximation to the binomial (see Chapter 6).

We could, of course, have used MINITAB, R, or JMP. We illustrate this with the following example using MINITAB and R:

Example 14.2.2 (Student's test scores data) *A chemistry professor from a large university wants to study the final paper scores obtained by the students in a general chemistry class. The following data gives the scores of 12 randomly selected students from that class:*

86 86 76 83 83 81 76 87 90 89 76 79

Test at the 5% level of significance the professor's hypothesis that the median score for the sampled population is 87.

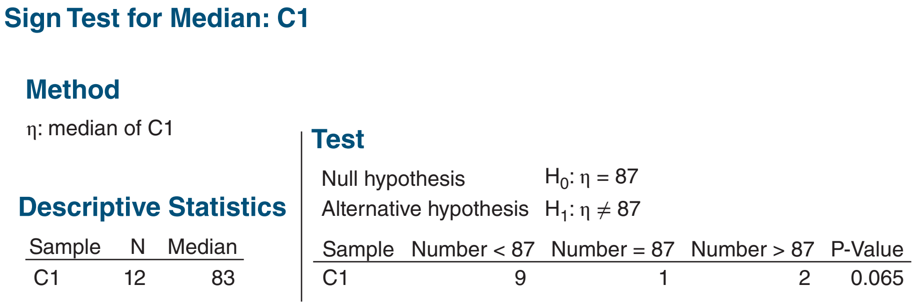
MINITAB

To test the hypothesis

$$H_0\colon\ M = 87 \quad \text{versus} \quad H_1\colon\ M \neq 87$$

using MINITAB, we proceed as follows:

1. Enter the data in column C1.
2. Select **Stat** > **Nonparametrics** > **1-Sample Sign**.
3. In the dialog box enter C1 in the box below Variables, and select **Test median** and enter 87, the specified value of the median under the null hypothesis. Then, select the appropriate **Alternative** hypothesis and click **OK**. The output that appears in the Session window is



```
Sign Test for Median: C1

Method
η: median of C1

Descriptive Statistics
Sample   N   Median
C1      12       83

Test
Null hypothesis          H0: η = 87
Alternative hypothesis   H1: η ≠ 87

Sample  Number < 87  Number = 87  Number > 87  P-Value
C1                9            1            2    0.065
```

Since the p-value is 0.065, which is greater than 0.05, we do not reject the null hypothesis that the median score for the sampled population is 87. We remark that MINITAB takes note of the fact that one observation with value equal to $M_0 = 87$ and automatically deletes it from consideration. For example, checking by hand, we find that

$$p\text{-value} = 2 \times \sum_{r=0}^{2} \binom{11}{r} \left(\frac{1}{2}\right)^{11} = 2 \times \{0.000488 + 0.005371 + 0.026855\} = 0.0654$$

as stated in the MINITAB output above.

USING R

The built-in function ‘SIGN.test()’ in R library(“BSDA”) can be used to conduct the required sign-test. The following R code can be used to obtain the test results for the data in Example 14.2.2.

```
x = c(86,86,76,83,83,81,76,87,90,89,76,79)
SIGN.test(x, md = 87, alternative = “two.sided”)

#R output
One-sample Sign-Test
data: x
s = 2, p-value = 0.06543
alternative hypothesis: true median is not equal to 87
95 percent confidence interval:
76.31909 86.89364
sample estimates:
median of x
83
```

As in the MINITAB procedure, the p-value is greater than 0.05. Therefore, we conclude that the median for the test scores, at an alpha level of 0.05, is not significantly different from 87.

14.2.2 The Wilcoxon Signed-Rank Test

Occasionally, we have a random sample that has been sampled from a population known to be *symmetric* about the unknown median, and the scale of measurement is at least *interval.* In such cases, the simple sign test discussed in Section 14.2.1 is not very desirable because it does not use all the information available in the sample. A more powerful test is the *Wilcoxon signed-rank test.* Since the sampled population is assumed to be symmetric, any conclusions made about the median are also valid for the mean.

The Wilcoxon signed-rank test proceeds by taking the differences between the measurements in the sample and a hypothesized location parameter, ranking the absolute magnitude of these differences, and then separating the ranks assigned to positive and negative differences. Any measurements that yield zero differences are discarded, and the sample size is adjusted by the number of measurements discarded.

Let $X_1, \ldots, X_n$ be a random sample from a population, symmetrically distributed, with an unknown median M. Then, we are interested in testing a hypothesis that the median takes some specified value. That is, the null and alternative hypotheses are, as in Section 14.2.1, namely:

$$H_0\colon\ M = m_0 \quad \text{versus} \quad H_1\colon\ M < m_0 \tag{14.2.5a}$$

$$H_0\colon\ M = m_0 \quad \text{versus} \quad H_1\colon\ M > m_0 \tag{14.2.5b}$$

$$H_0\colon\ M = m_0 \quad \text{versus} \quad H_1\colon\ M \neq m_0 \tag{14.2.5c}$$

To obtain the observed value of the test statistic, we take the following steps:

1. Find the differences between the measurements and the hypothesized median, that is, for each observation obtain
$$d_i = X_i - m_0$$
and record the absolute value, $|d_i|,\ i = 1, 2, \ldots, n$.
2. Rank the *absolute value* of all nonzero differences $|d_i|$ starting from the smallest to largest. If two or more $|d_i|$ are equal, then assign each one of them the average of the assigned ranks. This procedure is commonly called "breaking the ties." For example for two $|d_i|$ that are equal (i.e. they are tied and have ranks 7, 8) assign each the rank $(7 + 8)/2 = 7.5$.
3. Now find the sums of the ranks assigned to positive and negative differences separately, and denote them by T_+ and T_-, respectively.
4. We reject the null hypothesis in favor of the alternative whenever: for Equation (14.2.5a), T_+ is sufficiently small; for Equation (14.2.5b), T_- is sufficiently small; and for Equation (14.2.5c), Min(T_+, T_-) is sufficiently small. The critical values of the Wilcoxon signed-rank test are given in Table A.10.

We illustrate the Wilcoxon signed-rank test with the following example:

Example 14.2.3 (Bond strength of materials) *An article in* Annual Reviews of Material Research *2001 (p. 291) presents bond strengths for various energetic materials (explosives, propellants, and pyrotechnics). The bond strengths M for 15 such materials are shown below:*

323 312 300 284 283 261 207 183 180 179 174 167 167 157 120

(M = 220 is an industrial standard). Test at the 5% level of significance that the median bond strength has value 220.

Solution: The hypothesis that we would like to test is

$$H_0\colon\ M = 220 \quad \text{versus} \quad H_1\colon\ M \neq 220$$

The calculations for obtaining the observed value of the test statistic are shown in Table 14.2.1.

Table 14.2.1 Calculations for obtaining the value of the test statistic.

Obs.	d_i	$\|d_i\|$	Rank of $\|d_i\|$	Ranks after breaking ties	$T+$	$T-$
323	103	103	15	15.0	15.0	
312	92	92	13	13.0	13.0	
300	80	80	12	12.0	12.0	
284	64	64	11	11.0	11.0	
283	63	63	9	9.5	9.5	
261	41	41	4	4.5	4.5	
207	−13	13	1	1.0		1.0
183	−37	37	2	2.0		2.0
180	−40	40	3	3.0		3.0
179	−41	41	5	4.5		4.5
174	−46	46	6	6.0		6.0
167	−53	53	7	7.5		7.5
167	−53	53	8	7.5		7.5
157	−63	63	10	9.5		9.5
120	−100	100	14	14		14
Total					65.0	55.0

In this example, the test is two-sided, the observed test statistic is $Min(T_+, T_-) = Min(65, 55) = 55$. For the sample size 15, the value from Table A.10 indicates that we reject H_0 at significance level $\alpha = 0.05$ if the test statistic is less than or equal to 25. In this example, the test statistics has the value $55 > 25$, so we do not reject the null hypothesis. Based on these data, we may conclude that the median bond strength of various energetic materials is not different from 220.

Using MINITAB, the above test can be carried out as follows:

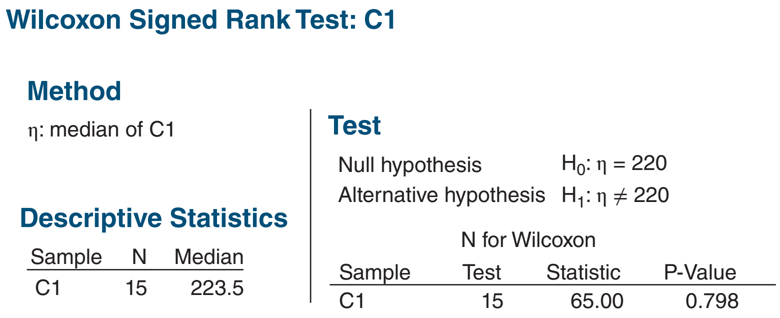
MINITAB

To test the hypothesis

$$H_0\colon\ M = 220 \quad \text{versus} \quad H_1\colon\ M \neq 220$$

using MINITAB, we proceed as follows:

1. Enter the data in column C1.
2. Select **Stat** > **Nonparametrics** > **1-Sample Wilcoxon ...** .
3. In the dialogue box **e**nter C1 in the box below Variables, and select **Test median** and enter 220, the specified value of the median under the null hypothesis. Then, select the appropriate **Alternative** hypothesis and click **OK**. The output that appears in the Session window is



Wilcoxon Signed Rank Test: C1

Method

η: median of C1

Descriptive Statistics

Sample	N	Median
C1	15	223.5

Test

Null hypothesis H_0: η = 220
Alternative hypothesis H_1: η ≠ 220

Sample	N for Test	Wilcoxon Statistic	P-Value
C1	15	65.00	0.798

Since the p-value is 0.798, which is greater than 0.05, we do not reject the null hypothesis that the median bond strength is 220.

USING R

The built-in function ‘wilcox.test()’ in R library(‘stats’) can be used to conduct the sign-test. The following R code can be used to obtain the required results for the data in Example 14.2.3.

```
x = c(323,312,300,284,283,261,207,183,180,179,174,167,167,157,120)
wilcox.test(x, alternative = "two.sided", mu=220)

#R output
Wilcoxon signed rank test with continuity correction
data: x
V = 65, p-value = 0.7982
alternative hypothesis: true location is not equal to 220
```

As in the MINITAB procedure, the p-value is greater than 0.05. Therefore, the previous conclusion stays the same.

14.2.3 Two-Sample Test

There are experimental situations, where we may think of taking n independent pairs of sample values, say $(X_1,\ X_1^{'}), (X_2,\ X_2^{'}), \ldots, (X_n,\ X_n^{'})$, from two populations having continuous cumulative distribution functions (c.d.f.’s) $F_1(x)$ and $F_2(x)$, where X_i has c.d.f. F_1 and $X_i^{'}$ has c.d.f. F_2. We can assume that each pair of values has measurements either on the same subject or subjects that have been matched with respect to certain criteria. Now consider a new variable U, where U is the difference of a pair of measurements so that $U_i = X_i - X_i^{'}$. Then, on the basis of the sample of n differences $U_1, \ldots, U_n$, we want to test the hypotheses

$$H_0\colon\ F_1 = F_2 \quad \text{versus} \quad H_1\colon\ F_1 > F_2 \tag{14.2.6a}$$

$$H_0\colon\ F_1 = F_2 \quad \text{versus} \quad H_1\colon\ F_1 < F_2 \tag{14.2.6b}$$

$$H_0\colon\ F_1 = F_2 \quad \text{versus} \quad H_1\colon\ F_1 \neq F_2 \tag{14.2.6c}$$

Now letting

$$Y_i = \begin{cases} 1, & \text{if } U_i > 0 \\ 0, & \text{if } U_i < 0 \end{cases}$$

We discard any pair that has $U_i = 0$, $i = 1, \ldots, n$, and then adjust the sample size n by reducing n by the number of discarded differences. For example if the starting sample size is 20, and three zero-differences are discarded, then the adjusted sample size is 17. Then, the test statistic is, for n, the adjusted sample size,

$$R = \sum_{i=1}^{n} Y_i$$

Now suppose that M_u, the median of the population represented by $U = X - X'$. If H_0: $M_u = 0$, is true, the medians of the two populations are the same, and we have

$$P(Y_i = 1) = P(X_i > X_i') = 1/2$$

since under H_0 the probability of the events $X_i > X_i^{'}$ and $X_i < X_i^{'}$ are equal. Furthermore, since the successive pairs are independent, under H_0, R is a random variable that has the binomial distribution $B(R = r|n, 1/2)$, that is,

$$P(R = r) = b(r) = \binom{n}{r} \left(\frac{1}{2}\right)^r \tag{14.2.7}$$

for $r = 0, 1, \ldots, n$. We say that the observed value r of R, is significantly large at the $100\alpha\%$ level of significance if $r \geq r_{L\alpha}$, where $r_{L\alpha}$ is the *smallest* integer for which

$$P(R \geq r_{L\alpha}) \leq \alpha \tag{14.2.8}$$

when H_0 is true. Hence, for a right-sided test, we would reject H_0 when using a one-side (right-sided) test at α level of significance if the observed $r \geq r_{L\alpha}$. Similarly, r is said to be significantly small at α level of significance if $R \leq r_{S\alpha}$, where $r_{S\alpha}$ is the *largest* integer for which

$$P(R \leq r_{S\alpha}) \leq \alpha \tag{14.2.9}$$

when H_0 is true. Thus, using a left-sided test of level α, we would reject H_0 if the observed $r \leq r_{S\alpha}$. Now let $r' = Min(r, n - r)$. We say that r' differs significantly from its expected value $n/2$ if r' does not fall in the interval $[r_{\alpha/2}, n - r_{\alpha/2}]$, where $r_{\alpha/2} < n/2$ is the largest integer for which

$$\begin{aligned} P(R < r_{\alpha/2}) = P(R > n - r_{\alpha/2}) \leq \alpha/2 \\ P(r_{\alpha/2} \leq R \leq n - r_{\alpha/2}) > 1 - \alpha \end{aligned} \tag{14.2.10}$$

when H_0 is true. Hence, using a two-sided test, we would reject H_0 at the α level of significance if the observed r' is such that $r' \leq r_{\alpha/2}$, or $r' \geq (n - r_{\alpha/2})$. Table 14.2.2 gives values of $r_{L\alpha}, r_{S\alpha}, r_{\alpha/2}$ for n up to 30 and for both $\alpha = 0.01$ and 0.05.

If n is larger than 30, we can approximate $r_{L\alpha}, r_{S\alpha}, r_{\alpha/2}$ by using the normal approximation (see Chapter 5) to the binomial, that is, approximately, for large n,

$$P(R \leq r_0) = \Phi\left(\frac{r_0 + \frac{1}{2} - \frac{n}{2}}{\frac{1}{2}\sqrt{n}}\right) \tag{14.2.11}$$

Table 14.2.2 Critical values of $r_{L\alpha}, r_{S\alpha}, r_{\alpha/2}$ when using the sign test.

	$r_{L\alpha}$		$r_{S\alpha}$		$r_{\alpha/2}$	
α n	0.01	0.05	0.01	0.05	0.01	0.05
5		5		0		
6		6		0		1
7	7	7	0	0	1	1
8	8	7	0	1	1	1
9	9	8	0	1	1	2
10	10	9	0	1	1	2
11	10	9	1	2	1	2
12	11	10	1	2	2	3
13	12	10	1	3	2	3
14	12	11	2	3	2	3
15	13	12	2	3	3	4
16	14	12	2	4	3	4
17	14	13	3	4	3	5
18	15	13	3	5	4	5
19	15	14	4	5	4	5
20	16	15	4	5	4	6
21	17	15	4	6	5	6
22	17	16	5	6	5	6
23	18	16	5	7	5	7
24	19	17	5	7	6	7
25	19	18	6	7	6	8
26	20	18	6	8	7	8
27	20	19	7	8	7	8
28	21	19	7	9	7	9
29	22	20	7	9	8	9
30	22	20	8	10	8	10

where $\Phi(z_\alpha)$ is the c.d.f. of the standard normal variable. Thus, if $\Phi(z_\alpha) = 1 - \alpha$, then we have

$$r_{L\alpha} \approx \frac{n+1}{2} + \frac{1}{2}\sqrt{n}\, z_\alpha \tag{14.2.12}$$

$$r_{S\alpha} \approx \frac{n-1}{2} - \frac{1}{2}\sqrt{n}\, z_\alpha \tag{14.2.13}$$

$$r_{\alpha/2} \approx \frac{n+1}{2} - \frac{1}{2}\sqrt{n}\, z_{\alpha/2} \tag{14.2.14}$$

Example 14.2.4 (Hemoglobin levels) *Table 14.2.3 (from Kenney and Keeping, 1956, vol. 1, p. 186) shows the hemoglobin (grams/100mL of blood) in 12 anemic rats before and after four weeks of added iron in the diet (0.5 milligram per day). Test, at the 1% and 5% level of significance, the hypothesis that the distribution of hemoglobin in anemic rats before and after the treatment is the same. We assume here that the interval of four weeks is sufficient to ensure that* X *(after) and X' (before) are independent.*

Table 14.2.3 Hemoglobin in rats before and after change in diet.

Rat number	X_i' (Before)	X_i (After)	$U_i = X_i - X_i'$	Y_i
1	3.4	4.9	1.5	1
2	3.0	2.3	−0.7	0
3	3.0	3.1	0.1	1
4	3.4	2.1	−1.3	0
5	3.7	2.6	−1.1	0
6	4.0	3.8	−0.2	0
7	2.9	5.8	2.9	1
8	2.9	7.9	5.0	1
9	3.1	3.6	0.5	1
10	2.8	4.1	1.3	1
11	2.8	3.8	1.0	1
12	2.4	3.3	0.9	1
				$\sum Y_i = 8$

Solution: Note that we observe $R = \sum_{i=1}^{12} Y_i$ to be $r = \sum_{i=1}^{12} Y_i = 8$, and consulting Table 14.2.2, we have under H_0 that $P(R \geq 10) \leq 0.05$, $P(R \geq 11) \leq 0.01$ and $P(R \leq 2) \leq 0.05$, $P(R \leq 1) \leq 0.01$. Furthermore, we have under H_0 that $P(2 \leq R \leq 10) > 0.99$ and $P(3 \leq R \leq 9) > 0.95$. Thus, no matter whether we consider a one- or a two-sided test, we do not have significance at either the 1% or 5% levels.

If we had assumed normality in this example and used the paired t-test (see Chapter 9), then similarly we would have found that the value of the test statistic is $t_{11} = 1.61$, which is not significant at either the 1% or 5% level.

Example 14.2.5 (Iron ore) *Thirty-six samples of ore were tested for their iron content by method A and method B. The data are shown in Table 14.2.4. Test the hypothesis that method A is equivalent to method B at the 5% level of significance.*

Solution: Suppose that X and X' are the iron contents determined by methods A and B, and take $U_i = X_i - X_i'$, $i = 1, \ldots, 36$. Further, let $Y_i = 1$ or 0 according to whether $U_i > 0$ or < 0, respectively, and let $R = \sum_{i=1}^{36} Y_i$; then a two-sided acceptance region for a test of the hypothesis that method A is equivalent to method B at the 5% level of significance is

$$[r_{0.025}, n - r_{0.025}]$$

Table 14.2.4 Data and calculations for obtaining the value of the test statistic.

X_i	$X_i^{'}$	U_i	Y_i	X_i	$X_i^{'}$	U_i	Y_i	X_i	$X_i^{'}$	U_i	Y_i
52	55	−3	0	52	51	1	1	54	50	4	1
55	54	1	1	55	55	0		53	54	−1	0
56	51	5	1	53	53	0		52	50	2	1
55	52	3	1	56	54	2	1	56	51	5	1
55	52	3	1	54	55	−1	0	55	50	5	1
52	50	2	1	52	51	1	1	56	55	1	1
56	54	2	1	52	51	1	1	56	50	6	1
55	53	2	1	53	55	−2	0	53	51	2	1
51	55	−4	0	53	53	0		55	54	1	1
54	54	0		56	52	4	1	51	55	−4	0
51	50	1	1	53	55	−2	0	52	53	−1	0
52	51	1	1	55	50	5	1	52	53	−1	0

But from Table 14.2.4, we find that the adjusted sample size is 32, because four pairs (X, X') have equal measurements and hence are discarded. Thus, $r_{0.025}$ is given by

$$\begin{aligned} r_{0.025} &\approx \frac{n+1}{2} - \frac{1}{2}\sqrt{n}\; z_{0.025} \\ &\approx \frac{33}{2} - \frac{1}{2}\sqrt{32}(1.96) = 10.95611 \end{aligned}$$

We use $r\ _{0.025} = 11$ and we have that $(n - r_{0.025}) = (32{-}11) = 21$. Hence, we should not reject the null hypothesis if $11 \leq r \leq 21$. In this example, we reject the null hypothesis because the value of r is 23, which does not fall in the defined acceptance region.

PRACTICE PROBLEMS FOR SECTION 14.2

1. Two tests are used to determine the hardness of a metal used in SUV bumpers. Ten samples of the metal under investigation are used for these tests. The hardness indexes under a certain scale produced by the two tests are shown below. Use the two-sample sign test to see whether the two tests produce equivalent results. Use $\alpha = 0.05$.

Sample	Test I	Test II
1	44	42
2	49	40
3	38	40
4	37	48
5	45	49
6	35	38

Sample	Test I	Test II
7	42	36
8	43	35
9	43	41
10	48	36

2. The following data give the heights in centimeters (cm) of 12 basketball players who were accepted with scholarships during the past 20 years. Use a one-sample sign test to test whether we can conclude that the median height of all the basketball players who were accepted with scholarships during that period is equal to 200 cm versus the hypothesis that the median height is greater than 200 cm. Use $\alpha = 0.05$. Determine the p-value for the test.

210	199	199	191	195	192	192	206	205	198	210	202

3. The following data give the placement test scores of a random sample of 20 candidates who have applied for undergraduate admission at a public university. Use the one-sample sign test to test that the median score in the placement test of all the applicants is greater than the required score of 115 points. Use $\alpha = 0.05$.

111	102	101	125	102	101	120	115	125	115
116	119	112	113	109	124	101	108	119	103

4. The following data give the cholesterol levels before and after the administration of a cholesterol-lowering drug in 10 randomly selected patients. Can we conclude at the 5% level of significance that the drug is ineffective?

Before:	183	174	145	140	177	153	175	173	140	152
After:	150	125	120	117	139	158	116	155	132	122

5. Reconsider the data in Problem 3 above. Use the Wilcoxon signed-rank test to test the hypothesis that the median score in the placement test of all the applicants is greater than the required score of 115 points. Use $\alpha = 0.05$.
6. Reconsider the data in Problem 2 above. Use the Wilcoxon signed-rank test to test whether we can conclude that the median height of all the basketball players who were accepted with scholarships during that period is equal to 200 cm versus the hypothesis that the median height is greater than 200 cm. Use $\alpha = 0.05$. Determine the p-value for the test.
7. The following data give the total cholesterol level (HDL + LDL + 20% of triglycerides) of 20 Americans between the ages of 30 and 40 years. Use the sign test to test the hypothesis at the 5% level of significance that the median cholesterol of American males in that age group is 150 mg/dL.

164	180	151	131	130	154	140	137	141	145
176	145	173	141	173	166	141	130	133	144

8. Repeat Problem 7 using the Wilcoxon signed-rank test.
9. The following data give “measured forced vital capacity” in eight asthma patients before and after a treatment. Use the Wilcoxon signed-rank test to test at the 5% level of significance that the treatment is effective.

Before:	3878	4011	3685	3384	4091	3451	3898	3098
After:	4249	3569	4262	3839	4177	4063	4304	3539

10. The following data give the time in minutes for 10 technicians to complete a project before and after an intensive training program. Use the Wilcoxon signed-rank test to test at the 1% level of significance that the training program is effective.

Before:	66	52	52	62	65	55	56	56	64	55
After:	47	48	60	43	52	40	58	54	49	52

14.3 MANN–WHITNEY (WILCOXON) W TEST FOR TWO SAMPLES

If two samples are not paired as in Section 14.2 and, in fact, if the samples are not necessarily of the same size, we may proceed as follows. Suppose that $(X_1, \ldots, X_m)$ and $(X_1', \ldots, X_n')$ are independent samples from populations having continuous c.d.f.s $F_1(x)$ and $F_2(x)$, respectively. We pool the two samples into a single sample of $m + n$ observations and let the order statistics of this sample be $Y_{(1)}, Y_{(2)}, \ldots, Y_{(m+n)}$.

Consider the ranks (subscripts) of all Y's that represent the elements of $(X_1, \ldots, X_m)$. Let the sum of these ranks be T, and let W be a random variable defined in terms of T as

$$W = mn + \frac{m(m+1)}{2} - T \tag{14.3.1}$$

Actually W is the number of the mn possible pairs (X_i, X_j') for which $X_i < X_j'$. W is called the *Mann–Whitney* statistic and T is called the *Wilcoxon* statistic. The test using T as test statistic is called the *Wilcoxon rank-sum* test.

It can be shown by rather complicated analysis (that we omit) that, if the hypothesis H_0: $F_1(x) \equiv F_2(x)$ is true; that is if both samples come from populations having identical c.d.f.s, then

$$E(W) = \frac{mn}{2} \tag{14.3.2}$$

$$Var(W) = \frac{mn(m+n+1)}{12} \tag{14.3.3}$$

It can be shown that as m and n both approach infinity, the random variable U, where

$$U = \frac{W - mn/2}{\sqrt{mn(m+n+1)/12}} \tag{14.3.4}$$

has, as its limiting distribution, the normal distribution $N(0,1)$. In practice, it has been found to be a good approximation that for m and n both greater than 8, Equation (14.3.4) is approximately distributed as $N(0,1)$.

Example 14.3.1 (Plutonium readings) *Two chemists, A and B, make 14 and 16 determinations of plutonium, respectively, with the results shown in Table 14.3.1. Numbers in parentheses are the rank of the observation in the combined sample. The problem is to determine whether the two chemists are doing equivalent work, or are obtaining significantly different results.*

Table 14.3.1 Data for Example 14.3.1.

Chemist-A		Chemist-B	
X	Ranks	X'	Ranks
263.36	(13)	286.53	(28)
254.68	(10)	254.54	(9)
248.64	(3)	284.55	(26)
272.68	(19)	253.75	(7)
261.10	(12)	283.85	(24)
287.33	(30)	252.01	(5)
268.41	(16)	245.26	(2)
287.26	(29)	275.08	(20)
276.32	(21)	286.30	(27)
243.64	(1)	272.52	(18)
256.42	(11)	282.90	(23)
282.65	(22)	266.08	(14)
250.97	(4)	267.53	(15)
284.27	(25)	252.05	(6)
		253.82	(8)
		269.81	(17)

Solution: Here $m = 14$, $n = 16$, and we want to test H_0: $F_A = F_B$ versus H_1: $F_A \neq F_B$. By combining the two samples into one sample and ordering the observations, we find that T, the sum of the ranks of the X observations in the combined sample, is 216, that is, $T = 13 + 10 + \cdots + 4 + 25 = 216$. Hence Equation (14.3.1) has the observed value

$$W = mn + \frac{m(m+1)}{2} - T = 224 + 105 - 216 = 113$$

We have from Equations (14.3.2) and (14.3.3) that if H_0 is true, then $E(W) = 112$ and $Var(W) = [(14)(16)(31)]/12 = 578.67$, so that the standard deviation of W is 24.06. Hence, the observed value of Equation (14.3.4) becomes

$$(113 - 112)/24.06 = 1/24.06 = 0.042 < z_{0.025} = 1.96$$

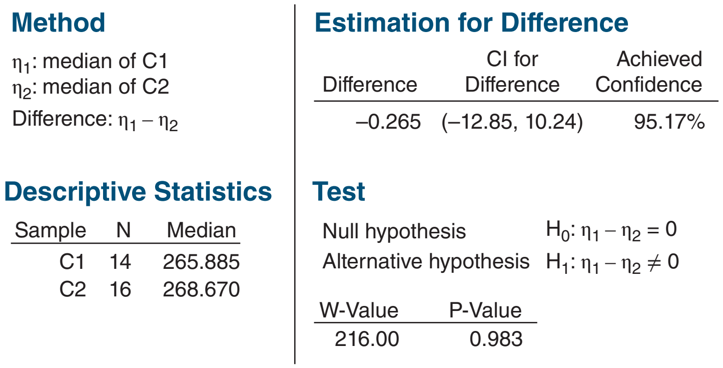
and for this two-sided test, the test statistic is not significant at the 5% level. Hence, we do not reject the null hypothesis that $F_A = F_B$; that is the chemists are doing equivalent work.

The problems in this section can be done by using one of the statistical packages discussed in this book.

MINITAB

1. Enter the data from two samples in columns C1 and C2, respectively.
2. From the Bar menu select **Stat** > **Nonparametric** > **Mann-Whitney**.
3. Enter C1 and C2 in the boxes next to **First sample** and **Second Sample**, respectively, and select the **confidence level**, that is, $1 - \alpha$. Then select the appropriate **Alternative** hypothesis (**in this problem not equal**) and click **OK**.

The output that appears in the Session window is given below. Note that the "W-value" (= 216.00) given in this MINITAB output indicates the sum of the ranks of the first (or X) sample in our data set and not the Mann-Whitney statistic.



```
Method                              Estimation for Difference

η1: median of C1                                      CI for        Achieved
η2: median of C2                     Difference     Difference      Confidence
Difference: η1 – η2                 ------------------------------------------
                                        –0.265   (–12.85, 10.24)      95.17%

Descriptive Statistics              Test

Sample    N    Median                Null hypothesis         H0: η1 – η2 = 0
-----------------------              Alternative hypothesis  H1: η1 – η2 ≠ 0
    C1   14   265.885
    C2   16   268.670                W-Value       P-Value
                                    -----------------------
                                      216.00        0.983
```

Since the p-value is greater than 0.05, we do not reject the null hypothesis that $F_A = F_B$ the chemists are doing equivalent work.

USING R

Similar to Example 14.2.3, the built-in function 'wilcox.test()' in R library "stats" can be used to conduct the Mann-Whitney test. The following R code can be used to obtain the required results for the data in Example 14.3.1. Note that in order to get the same value for the Mann-Whitney test statistic defined in Equation (14.3.1), we have to switch the order of the sample arguments in ‘wilcox.test()’ function as shown below.

```
x = c(263.36,254.68,248.64,272.68,261.1,287.33,268.41,287.26,
276.32,243.64,256.42,282.65,250.97,284.27)
y = c(286.53,254.54,284.55,253.75,283.85,252.01,245.26,275.08,
286.3,272.52,282.9,266.08,267.53,252.05,253.82,269.81)

wilcox.test(y, x, alternative = “two.sided”)

#R output
Wilcoxon rank sum test
data: y and x
W = 113, p-value = 0.9837
alternative hypothesis: true location shift is not equal to 0
```

As in the MINITAB procedure, the p-value is greater than 0.05. Therefore, the previous conclusion stays the same.

PRACTICE PROBLEMS FOR SECTION 14.3

1. A computer manufacturing company purchases memory chips from two different suppliers. The following data give the thickness of the coated film (coded data) on eight randomly selected chips received from two suppliers. Can we conclude at the 5% level of significance that thickness of coated films on chips shipped by the two suppliers is the same, using the Mann-Whitney test?

Supplier I:	29	30	26	33	33	26	23	20
Supplier II:	26	18	15	15	21	30	24	30

2. A coordinator of an engineering program was interested in testing the effectiveness of onsite and online courses. Twenty-three students were selected to participate in this research project. Ten of them were assigned to attend the onsite class and the remaining 13 were assigned to attend the online class. At the end of a seven-week session, all the students were given the same test. The data below give the scores of all the students who participated in this project. Based on these data, can we conclude at the 5% level of significance that the two methods of teaching are equally effective using the Mann-Whitney test?

Onsite:	96	84	85	100	86	80	83	93	100	83			
Online:	88	83	81	76	85	94	79	95	79	85	90	90	96

3. The following data give the time (in eight-hour shifts) taken to complete a project by 10 technicians randomly selected from each of the two plants of a company. Based

on these data, can we conclude at the 5% level of significance that the standards of hiring technicians at the two plants are different using the Mann-Whitney test?

Plant I:	27	25	27	26	20	28	25	20	23	22
Plant II:	21	25	21	20	19	22	20	21	24	18

4. A medical doctor wished to test the effect of a cholesterol-lowering drug when it is prescribed to children in one of two forms, tablet or suspension. The following data give the reduction in cholesterol levels (in mg/dL) after a full four weeks of treatment. Can we conclude at the 5% level of significance that both forms of the drug have the same effect in lowering the cholesterol level using the Mann-Whitney test?

Tablet:	37	38	33	38	39	34	40	31		
Suspension:	43	43	35	47	45	45	46	39	45	46

5. The following data give drying times (in hours) of two brands of oil-based paint. Use the Mann-Whitney test to test the hypothesis at the 5% level of significance that both paint brands dry in the same number of hours.

Brand 1:	8.5	9.0	8.1	10.0	9.3	9.0	8.0	9.2
Brand 2:	10.3	8.5	8.9	10.2	9.8	8.5	8.9	9.3

6. The following data give the total yields of a chemical produced by using the same catalyst at two different temperatures. Use the Mann-Whitney test to test the hypothesis at the 5% level of significance that the two temperatures result in the same yield.

Temperature 1:	61	60	78	65	68	64		
Temperature 2:	67	76	72	73	82	79	71	81

14.4 RUNS TEST

14.4.1 Runs above and below the Median

Suppose that we take a sample of $2n$ elements $(X_1, X_2, \ldots, X_{2n})$ from a *continuous* c.d.f. $F(x)$, and consider the problem of examining the fluctuations in the sequence of the $2n$ drawings for evidence of nonrandomness. One simple way of doing this is as follows: graph the X_i against i, $i = 1, \ldots, 2n$, as shown in Figure 14.4.1.

If the order statistics of the sample are $(X_{(1)}, X_{(2)}, \ldots, X_{(2n)})$, and since the probability that two or more of the X's are equal in the sample is equal to zero, there will be a gap

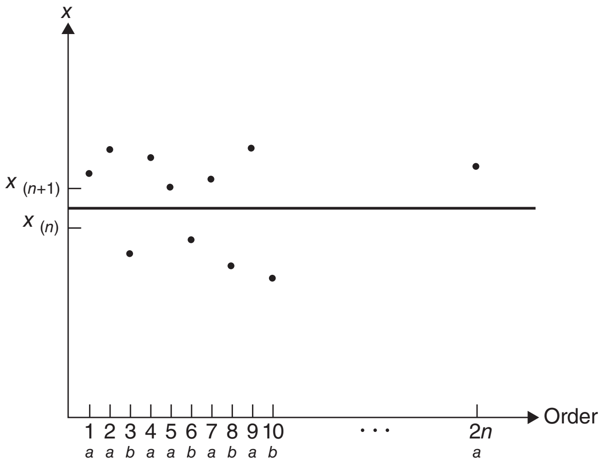


Figure 14.4.1 Chart for determining runs above and below the median.

between $X_{(n)}$ and $X_{(n+1)}$. To complete the plotting, draw a horizontal line half-way between $X_{(n)}$ and $X_{(n+1)}$, as in Figure 14.4.1.

There will be n points above the line and n below. For each X_i above the line, write a, and for each point below the line, write b, as in Figure 14.4.1, so that there are n, a symbols and n, b symbols. The total number of distinct possible arrangements of a's and b's is $\binom{2n}{n}$. If the sample is a random sample from any continuous c.d.f. $F(x)$, then these $\binom{2n}{n}$ different possible arrangements have equal probabilities. For any given arrangement, there are clusters of one or more a's separated by clusters of one or more b's.

The *total number of clusters of a's and b's*, say U, is called the *number of runs above or below the median line.* Note that if there had been an odd number of elements in the sample, say $2n + 1$, we could have drawn the horizontal line through the order statistic $X_{(n+1)}$, that is, the median of the sample $(X_1, \ldots, X_{2n+1})$. We would then have n points above the median and n below. If the measurements in the sequence $X_1, \ldots, X_{2n}$, which is the order in which the measurements were drawn, exhibit wide swings or jumps from one general level to another, then the sequence tends to make U significantly small. If the measurements tend to alternate too much, U tends to be significantly large. In either case, U is a reasonable indicator of nonrandomness.

It can be shown that the probability function $p(u)$ of U, assuming that $(X_1, \ldots, X_{2n})$ is a *random sample* from a continuous c.d.f., is given by

$$p(u) = 2\frac{\binom{n-1}{\frac{u}{2}-1}^2}{\binom{2n}{n}} \qquad \text{if } u = 2, 4, \ldots, 2n \tag{14.4.1}$$

$$p(u) = 2\frac{\binom{n-1}{\frac{u}{2}-\frac{1}{2}}\binom{n-1}{\frac{u}{2}-\frac{3}{2}}}{\binom{2n}{n}} \qquad \text{if } u = 1, 3, \ldots, 2n-1 \tag{14.4.1a}$$

The mean and variance of U are given by

$$\mu_u = n + 1 \tag{14.4.2}$$

$$\sigma_u^2 = \frac{n(n-1)}{2n-1} \tag{14.4.3}$$

(For proofs of results in Equations (14.4.1), (14.4.1a), (14.4.2), and (14.4.3), see Wilks (1962).) It is known that for large n (> 10), U has approximately a normal distribution with the mean and variance given in Equations (14.4.2) and (14.4.3). As mentioned previously, we will suspect nonrandomness if U is either too small or too large. For $n \geq 10$, we then reject randomness if $|U - \mu_u| \geq \sigma_u z_{\alpha/2}$, at significance level α.

Example 14.4.1 (Dimension of rheostat knobs) *The data in Table 14.4.1 give sample measurements on a certain dimension of rheostat knobs. Examine the sequence of measurements in Table 14.4.1 for evidence of nonrandomness with respect to runs.*

Solution: Determining the order statistics of the observations, we find that

$$X_{(12)} = 0.1412 \text{ in.} \quad \text{and} \quad X_{(13)} = 0.1414 \text{ in.}$$

We take, then, as the central line for a test of randomness $X = (X_{(12)} + X_{(13)})/2 = 0.1413$ in. Writing a if an observation has value > 0.1413 and b if it is < 0.1412, we obtain the sequence

$$b \quad a \quad a \quad b \quad a \quad b \quad b \quad a \quad b \quad a \quad a \quad a \quad b \quad a \quad b \quad a \quad b \quad a \quad b \quad b \quad a \quad b \quad a \quad b$$

Here the number of runs (clusters) U is observed to be $U = 19$ with $n = 12$, $E(U) = 13$, $\sigma_u^2 = (12)(11)/23 = 5.74$, so $\sigma_u \approx 2.4$.

Table 14.4.1 Data on rheostats.

Observation		Rank order	Observation		Rank order
number	Observed	of magnitude	number	Observed	of magnitude
1	0.1367	2	13	0.1394	1
2	0.1414	13	14	0.1422	19
3	0.1415	15	15	0.1412	12
4	0.1406	7	16	0.1417	17
5	0.1416	16	17	0.1402	5
6	0.1404	6	18	0.1425	20
7	0.1400	4	19	0.1407	8
8	0.1418	18	20	0.1408	9
9	0.1410	11	21	0.1430	22
10	0.1432	23	22	0.1398	3
11	0.1448	24	23	0.1415	14
12	0.1428	21	24	0.1409	10

A two-tailed test for the hypothesis of randomness at the 1% level of significance for this example is given by the following rule: reject the hypothesis if $|U - 13| \geq (2.4)(2.575) = 6.18$; otherwise, do not reject the hypothesis. In this case, it is observed that $|U - 13| = |19 - 13| = 6$, so we do not reject the hypothesis and conclude that the sequence is random, that is, does not exhibit significant nonrandomness.

The general procedure for testing whether an observed value of U is significantly large or small or differs significantly from its mean value of $(n + 1)$ one way or the other (two-tailed test) is similar to that used in the sign test except, of course, that we use the probability function of U instead of R of Equation (14.2.2). In most practical situations, nonrandomness tends to reveal itself in significantly small values of U.

14.4.2 The Wald–Wolfowitz Run Test

The ideas of Section 14.4.1 extend to the case in which we have two samples, say $X_1, \ldots X_{n_1}$ and $X_1^{'}, \ldots, X_{n_2}^{'}$, from two populations having continuous c.d.f.'s $F_1(x)$ and $F_2(x)$, respectively. Suppose now that we intend to test

$$\begin{array}{ll} H_0\colon\ F_1(x) \equiv F_2(x) & \text{for all } x \\ \quad\text{versus} & \\ H_1\colon\ F_1(x) \neq F_2(x) & \text{for at least one } x \end{array}$$

The procedure is as follows: We combine the two samples into one pooled sample and then order these observations. For example, we might obtain a sequence such as

$$X_{(1)}, X_{(2)}, X_{(1)}^{'}, X_{(2)}^{'}, X_{(3)}, X_{(3)}^{'}, X_{(4)}^{'}, \ldots$$

where $X_{(i)}(i = 1, 2, \ldots, n_1)$ is the ith-order statistic of the X-sample and $X_{(j)}^{'}(j = 1, 2, \ldots, n_2)$ is the jth-order statistic of the X'-sample.

We now replace each observation by a 0 or 1, according to whether we encounter an X or an X'. The sequence above, for example would be

$$0011011\ldots$$

A cluster of one or more zeros, or similarly, a cluster of one or more ones, is called a run.

Now there are $\binom{n_1+n_2}{n_1}$ possible distinguishable arrangements or permutations of the n_1 0's and n_2 1's. Under the null hypothesis that $F_1(x) \equiv F_2(x)$, these arrangements are equally likely. For each permutation, there will be a total number U of runs, and we can use this to test the hypothesis H_0 that $F_1(x) \equiv F_2(x)$.

The probability function of U is given by

$$\begin{aligned} p(u) &= 2\frac{\binom{n_1-1}{\frac{u}{2}-1}\binom{n_2-1}{\frac{u}{2}-1}}{\binom{n_1+n_2}{n_1}} \quad \text{if } u \text{ is even} \\ p(u) &= \frac{\binom{n_1-1}{\frac{1}{2}(u-1)}\binom{n_2-1}{\frac{1}{2}(u-3)} + \binom{n_1-1}{\frac{1}{2}(u-3)}\binom{n_2-1}{\frac{1}{2}(u-1)}}{\binom{n_1+n_2}{n_1}} \quad \text{if } u \text{ is odd} \end{aligned} \tag{14.4.4}$$

Under H_0 it can be shown that mean and variance of U are as follows:

$$\mu_u = E(U) = \frac{2n_1n_2}{n_1+n_2} + 1 \tag{14.4.5}$$

$$\sigma_u^2 = Var(U) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \tag{14.4.6}$$

Furthermore, if n_1 and n_2 are large, U has approximately a normal distribution with mean and variance given in Equations (14.4.5) and (14.4.6), respectively. The approximation of the distribution of U by the normal is usually adequate for practical purposes when both n_1 and n_2 exceed 10. The reader should note that the results (14.4.4), (14.4.5), and (14.4.6) reduce to Equations (14.4.1), (14.4.2), and (14.4.3), respectively, if $n_1 = n_2 = n$.

Example 14.4.2 (Testing two samples using a run test) *The following two samples of measurements were obtained from sampling two populations, I and II. The problem is to test, using the theory of runs, the hypothesis that the two samples are from populations having identical c.d.f.s.*

Population I (X):	*25*	*30*	*28*	*34*	*24*	*25*	*13*	*32*	*24*	*30*	*31*	*35*			
Population II (X′):	*44*	*34*	*22*	*8*	*47*	*31*	*40*	*30*	*32*	*35*	*18*	*21*	*35*	*29*	*22*

Solution: Here $n_1 = 12, n_2 = 15$, so $\mu_u = 14.3$ and $\sigma_u^2 = 6.32$, and the standard deviation of U is $\sigma_u = 2.51$. The combined sample when ordered gives rise to the sequence

$$8_{x'}, 13_x, 18_{x'}, 21_{x'}, 22_{x'}, 22_{x'}, 24_x, 24_x, 25_x, 25_x, 28_x, 29_{x'}, 30_x, 30_x,$$
$$30_{x'}, 31_x, 31_{x'}, 32_x, 32_{x'}, 34_x, 34_{x'}, 35_x, 35_{x'}, 35_x, 40_{x'}, 44_{x'}, 47_{x'},$$

where the x and x' subscripts denote measurements from the samples of X and X', respectively. We denote an X observation by 0 and an X' observation by 1. Note that in the data we have encountered repeated elements within and across samples for the values 30, 31, 32, 34, and 35. In fact $30_x, 30_x$, and $30_{x'}$ occur at positions 13, 14, and 15 in the combined ordered sample above, and so on. One widely used procedure to break the ties is illustrated by the following example. Toss a coin. If the toss results in a head, we replace 30_x by a zero, $30_{x'}$ by a one; if tails occurs, we replace 30_x by a one and $30_{x'}$ by a zero. In this example, performing this randomization procedure might lead to the following sequence:

1 0 1 1 1 1 0 0 0 0 0 1 0 1 0 1 0 1 0 1 0 1 0 1 1 1 1

which gives $U = 17$ runs. Now we reject the null hypothesis using the normal approximation at the 1% level of significance if $|U - \mu_u| \geq \sigma_u z_{\alpha/2}$, which for this example with $\alpha = 0.01$, is.

Reject H_0: $F_1 = F_2$ if

$$|U - 14.3| \geq 2.51(2.575) \approx 6.5$$

and do not reject otherwise. But the observed value of U is 17, so

$$|U - 14.3| = |17 - 14.3| = 2.7 < 6.5$$

Hence we do not reject the hypothesis that the two samples come from populations having identical c.d.f.'s.

PRACTICE PROBLEMS FOR SECTION 14.4

1. The following sequence shows the wins (W) and losses (L) of a baseball team for 40 consecutive games in a given season (there are usually 162 games in a season). Use the runs test to test a hypothesis that the sequence of wins and losses is random. Use $\alpha = 0.05$.

   W W W L L W L L L W W W W L LW W W L L W L L L W W W L L W L W

   W W L L W L L W

2. Use the Wald–Wolfowitz run test for data in Problem 2 of Section 14.3. Can we conclude, based on these data, that the two types of courses have different distributions? Use $\alpha = 0.05$.
3. A quality control engineer of a manufacturing company classifies a manufactured part as conforming (C) or nonconforming (N) depending on whether or not it meets the specifications. The company received a new machine and the engineer inspects 45 consecutive parts produced by the new machine for their conformity. The parts produced turned out to be as shown below. Use the runs test to test that the sequence of conforming and nonconforming is random. Use $\alpha = 0.01$.

   C C C N C C C N N C C C C C N C C C C C C C N N C C C N C C

   N C C C C N C C C C C C N N C

4. The following data give the number of school years spent by some American adults of ages between 30 and 40 years who were randomly selected from two different socioeconomic groups. Using the Wald–Wolfowitz run test, can we conclude based on the data below that the two populations have different distributions? Use $\alpha = 0.05$.

Group I:	14	14	13	12	13	9	8	15	10	8	12	13			
Group II:	16	14	14	16	17	22	16	15	22	22	12	18	17	21	20

5. The following data give the gender of the first 25 students who were admitted to a medical school. Use the runs test to test that the sequence of males and females is random. Use $\alpha = 0.01$.

   M M F F F M F F M M M F F M F M M F F F M M F M M

6. The following data give the departures from the desired specified tensile strength for 24 pieces of copper wires. Find whether we can conclude that the pattern of departure above and below the specified value is the result of a nonrandom process (read the data in row 1 first and then proceed to row 2). Use $\alpha = 0.05$.

23	−24	30	28	−30	2	56	−30	−49	−23	33	42
34	20	−41	26	−16	−18	37	24	−44	−29	28	−30

7. The following data give the number of interviews for a particular residency received by some medical students from two different medical schools. Using the Wald–Wolfowitz run test, can we conclude, from these data, that the two populations have different distributions? Use $\alpha = 0.05$ (note that $n_A = 11, n_B = 12$).

School A:	9	8	9	10	9	9	9	9	8	10	10	
School B:	8	9	7	7	9	9	7	9	7	9	8	9

8. A dietician wants to investigate the effect of a dieting program on males and females. The following data give the weight loss (in lbs) of a group of males and females after they were on that program for six weeks. Using the Wald–Wolfowitz run test, can we conclude from these data that the two populations have different distributions? Use $\alpha = 0.05$.

Females:	18	24	18	22	20	21	25	18	22	20	20	
Males:	19	15	10	16	10	14	15	13	15	10	11	20

14.5 SPEARMAN RANK CORRELATION

Let $(X_1, X_1^{'}), (X_2, X_2^{'}), \dots, (X_n, X_n^{'})$ be a random sample from a continuous bivariate population, that is, each pair of observations $(X_i, X_i^{'}), i = 1, 2, \dots, n$, represents a pair of measurements on the same subject and the scale of measurement is at least ordinal. Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ and $X_{(1)}^{'}, X_{(2)}^{'}, \dots, X_{(n)}^{'}$ be the order statistics of the X and X' measurements, respectively. We now rank X and X' measurements of the sample $(X_1, X_1^{'}), (X_2, X_2^{'}), \dots, (X_n, X_n^{'})$ so that if $X_i = X_{(1)}$, then $R(X_i) = 1$, and if $X_i = X_{(n)}$, then $R(X_i) = n$, and so on. Similarly, if $X_i^{'} = X_{(1)}^{'}$, then $R(X_i^{'}) = 1$, and if $X_i^{'} = X_{(n)}^{'}$, then $R(X_1^{'}) = n$. If some X and/or X' measurements are repeated, then they are assigned ranks equal to the average of their ranks. For example, if an X measurement that is 8 (say) is repeated three times and the actual ranks within the set of X_i measurements are 5, 6, and 7, then each measurement 8 is assigned a rank of 6, determined from $(5 + 6 + 7)/3 = 6$. This process of reassigning the ranks is usually known as a process of breaking a tie. Having done that, we may now want to test at the α level of significance one of the following hypotheses.

H_0: X and X' are not correlated,

versus

(a) H_1: X and X' are positively or negatively correlated
(b) H_1: X and X' are positively correlated
(c) H_1: X and X' are negatively correlated

Hypothesis (a) is two-sided, whereas hypotheses (b) and (c) are one-sided. The test statistic for testing H_0 against any of the hypotheses (a), (b), or (c) is known as the Spearman rank correlation and given by

$$r_s = 1 - 6\sum_{i=1}^{n} \frac{d_i^2}{n(n^2-1)}, \quad d_i = R(X_i) - R(X_i^{'}), \qquad i = 1, 2, \ldots, n \qquad (14.5.1)$$

Note that r_s satisfies the inequality $-1 \leq r_s \leq 1$. Furthermore, the reader can easily verify that

$$r_s = 1 \quad \text{if } R(X_i) = R(X_i^{'}) \qquad (14.5.2)$$

$$r_s = -1 \quad \text{if } R(X_i) = R(X_{n+1-i}^{'}) \qquad (14.5.3)$$

for $i = 1, 2, \ldots, n$. We say that the ranking of X and X' are in perfect agreement when $r_s = 1$ and in perfect disagreement when $r_s = -1$. In other words, when $r_s = 1$, there is direct perfect association, and when $r_s = -1$, there is *perfect* disagreement between X and X'. This means generally that large values of $|r_s|$ tend to support the alternative hypotheses. The critical values of r_s are given in Table A.16. If the absolute observed value of r_s is greater than the tabled value, we reject the null hypothesis. Otherwise, we do not reject the null hypothesis. We illustrate this test with the following example.

Example 14.5.1 (Job applicants' interview scores) *A large engineering company received 10 applications from well-qualified engineers for the position of senior engineer. These applicants are interviewed by two senior managers who independently awarded them interview scores. The scores are shown in Table 14.5.1. Do the data provide sufficient evidence at the 5% level of significance of an association between the scores awarded by the two interviewers?*

Solution: In this example, the hypothesis that we wish to test is

H_0: X and X' are independent
versus
H_1: X and X' are positively or negatively correlated

Table 14.5.1 Interview scores and their ranks.

X_i	$X_i^{'}$	$R(X_i)$	$R(X_i^{'})$	d_i	d_i^2
87	82	2.0	3.0	−1.0	1.00
90	83	5.0	4.5	0.5	0.25
95	84	9.5	7.0	2.5	6.25
94	91	8.0	10.0	−2.0	4.00
85	80	1.0	1.5	−0.5	0.25
92	83	6.5	4.5	2.0	4.00
92	90	6.5	9.0	−2.5	6.25
89	80	4.0	1.5	2.5	6.25
95	84	9.5	7.0	2.5	6.25
88	84	3.0	7.0	−4.0	16.00
					$\sum d_i^2 = 20.5$

We have $n = 10$ and $\sum_{i=1}^{10} d_i^2 = 1 + 0.25 + \cdots + 16 = 50.5$. Hence, the test statistic, using Equation (14.5.1), is observed to be

$$r_s = 1 - \frac{6(50.5)}{10(100-1)} = 0.693$$

Table A.16 indicates that for $n = 10$, for the two-sided test at the 5% level of significance, the critical value of the test statistics r_s is 0.6364, which is smaller than the observed value of r_s. Hence, we reject the null hypothesis at significance level 0.05 that the scores awarded by two interviewers are independent.

For large n (> 30), we can use a Z statistic that is distributed approximately as the standard normal $N(0, 1)$, where Z is defined as

$$Z = r_s\sqrt{n-1} \tag{14.5.4}$$

We will discuss other nonparametric methods, namely, the Kruskal–Wallis test for one-way ANOVA and the Friedman test for two-way ANOVA, in Chapter 17.

PRACTICE PROBLEMS FOR SECTION 14.5

1. The following data give the scxores in the final exams on differential equations and quantum mechanics that were received by 10 students randomly selected from a freshmen engineering class. Do these data provide sufficient evidence to indicate an association between the two scores? Use $\alpha = 0.05$.

DE:	87	92	93	94	97	86	98	95	88	96
QM:	82	91	81	88	83	86	82	88	91	81

2. Ten female candidates are ranked in a beauty competition by two judges. These ranks are shown below. Do these data provide sufficient evidence to indicate an association between the two ranks? Use $\alpha = 0.05$.

Judge I:	4	7	9	5	10	2	8	3	1	6
Judge II:	2	10	3	6	5	1	7	8	9	4

3. The following data show heights in centimeters (cm) and average scores per game of 12 randomly selected basketball players. Do these data provide sufficient evidence to indicate an association between the heights and the average scores per game? Use $\alpha = 0.10$.

Heights:	201	199	194	188	198	201	202	197	192	195	205	198
Scores:	16	22	25	22	21	23	23	22	19	23	17	15

4. In a study of the relationship between age and LDL (mg/dL), data were obtained on 10 subjects between 40 and 70 years. The following data give the ages and LDL for each of the 10 subjects. The experimenter wants to know if we can conclude at the 5% level of significance that age and LDL are positively correlated.

Age:	58	66	47	52	53	51	54	58	52	56
LDL:	131	162	144	145	130	115	118	151	122	157

5. The following data give the IQ level and the scores obtained in a standardized test for 12 candidates. Determine the Spearman rank correlation and test at the 5% level of significance that there is a positive correlation between the IQ level and the test scores.

IQ:	92	117	100	90	130	108	121	130	105	123	129	114
Test:	79	88	79	86	86	84	77	95	82	95	84	95

6. A manager of a manufacturing company wished to study the years of service and productivity of a group of technicians. The following data give the years of service and the productivity for 10 technicians. Determine the Spearman rank correlation and test at the 5% level of significance that there is a correlation between the years of service and productivity.

Years of service:	11	13	19	12	13	12	16	18	16	15
Productivity:	113	119	107	117	114	109	115	120	117	118

7. The following data give systolic blood pressure readings taken by a doctor and her nurse on the last 10 patients who visited the doctor's office. Determine the Spearman rank correlation and test at the 5% level of significance that there is a positive correlation between the systolic blood pressure taken by the doctor and her nurse.

Doctor:	122	130	115	130	124	125	130	129	126	124
Nurse:	135	130	140	135	128	137	130	125	123	136

8. The following data give the white blood cell counts and the duration of follow-up (in months) after the operations of some cancer patients. Determine the Spearman rank correlation and test at the 5% level of significance that there is a negative correlation between the white blood cell counts and the duration of follow-up.

Duration of follow-up:	15	11	12	19	18	19	13	15
White blood cell count:	2773	3207	3535	3512	2792	3143	3302	3153

14.6 USING JMP

This section is not included in the book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Fifteen diabetic patients are given 1000 mg of Metformin (500 mg twice a day) and two weeks later their serum sugar levels were as shown below. Do the data give sufficient evidence to indicate that the patients on Metformin have median serum sugar level of 140 mg/dL? Use the Sign test at the 5% level of significance.

134	149	141	143	126	130	147	134
134	144	128	149	135	126	149	

2. A consumer group rated 12 manufactured products imported from a single source on a scale of 1–10. The data are shown below. Do the data provide sufficient evidence to indicate that the median score is greater than 5? Use the sign test at the 1% level of significance.

2	9	6	7	8	5	10	2	8	5	3	4

3. The following data show measurements of the corrosion effects of various soils for coated and uncoated steel pipe (from Hoel, 1954). Use the sign test to test the hypothesis that the particular coating used has no effect on corrosion. Use $\alpha = 0.05$.

Uncoated, x:	42	37	61	74	55	57	44	55	37	70
	52	55	60	48	52	44	56	44	38	47
Coated, x':	39	43	43	52	52	59	40	45	47	62
	40	27	50	33	56	36	54	32	39	40

4. Use the sign test for Problem 24 in Review Practice Problems in Chapter 9.

5. Two samples of 30 observations each from populations A and B are such that when the Mann-Whitney W test for two samples is used, the sum of the ranks in the pooled sample of the X' measurements (from population B) is found to be 1085. Hence, the Mann–Whitney test statistic has value $W = nm + (m(m+1))/2 - T = 280$. Is the value of W significantly small at the 1% level?

6. Use the Wald–Wolfowitz run test for the data of Problem 9 of Review Practice Problems in Chapter 8 to test the null hypothesis that the samples come from populations having identical continuous c.d.f.s $F_1(x)$ and $F_2(x)$. Use $\alpha = 0.05$.

7. Twenty-four samples of four insecticide dispensers were taken periodically during a production period. The average charge weights (in grams) of the 24 samples are shown below. Use the median test on the sample means to test for randomness. Use $\alpha = 0.05$.

Sample no.	$\bar{X}$	Sample no.	$\bar{X}$
1	471.5	13	457.5
2	462.2	14	431.0
3	458.5	15	454.2
4	476.5	16	474.5
5	461.8	17	475.8
6	462.8	18	455.8
7	464.0	19	497.8
8	461.0	20	448.5
9	450.2	21	453.0
10	479.0	22	469.8
11	452.8	23	474.0
12	467.2	24	461.0

8. The total thicknesses X of the four pads of 36 half-ring (H-R) mounts for aircraft engines, taken periodically from the production line, were found to be as shown below. Determine whether the total number of runs above and below the median of this sequence of values of X is significantly different from its expected value at the 5% level of significance, under the hypothesis that X is under statistical control.

H-R#	X	H-R#	X	H-R #	X
1	1.5936	13	1.5900	25	1.5920
2	1.5906	14	1.5925	26	1.5918
3	1.5982	15	1.5924	27	1.5908
4	1.5901	16	1.5901	28	1.5913
5	1.5869	17	1.5902	29	1.5922
6	1.5898	18	1.5904	30	1.5915
7	1.5930	19	1.5910	31	1.5927
8	1.5894	20	1.5926	32	1.5913
9	1.5885	21	1.5890	33	1.5911
10	1.5905	22	1.5895	34	1.5903
11	1.5904	23	1.5936	35	1.5916
12	1.5902	24	1.5933	36	1.5939

9. A sequence of 100 measurements on a process, supposedly under statistical control, was analyzed for runs above and below the median. The total number of runs above

and below the median was observed to be 39. Is this a significantly low value at the 5% level of significance? Show your calculations.

10. During a 45-day production period in a cement plant, test cubes were taken each day, and the compressive strengths (in kg/cm^2) of the test cubes were determined with the following results (data from Hald, 1952). Test the hypothesis of randomness at the 5% level of significance in this sequence by using the run test of above and below the median.

440	433	475	418	462	433	438	469	465	500	524	491	479	442	463	423
447	452	498	497	473	485	488	440	442	451	423	452	475	526	490	511
492	509	472	428	456	417	452	468	475	502	507	477	445			

11. The 58 observations shown below were obtained by Millikan (1930) for the charge on an electron in 10^{-10} esu (CGS) units. By using runs above and below the median, test the hypothesis that the variation in this sequence is behaving randomly. Specify the value of α that you are using.

4.781	4.771	4.768	4.788	4.790
4.795	4.789	4.801	4.783	4.747
4.769	4.772	4.785	4.740	4.769
4.792	4.789	4.783	4.775	4.806
4.779	4.764	4.808	4.761	4.779
4.775	4.774	4.771	4.792	4.785
4.772	4.778	4.809	4.758	4.790
4.791	4.791	4.790	4.764	4.777
4.782	4.777	4.779	4.810	4.749
4.769	4.765	4.788	4.799	4.781
4.764	4.785	4.772	4.779	
4.776	4.805	4.791	4.797	

12. Two analysts took repeated readings on the hardness of city water. The data are shown below. Determine whether one analyst has a tendency to read the instruments differently from the other, using the Mann–Whitney Wilcoxon test (data from Bowker and Lieberman, 1959). Use $\alpha = 0.05$.

Analyst A:	0.46	0.62	0.37	0.4	0.44	0.58	0.48	0.53		
Analyst B:	0.82	0.61	0.89	0.51	0.33	0.48	0.23	0.25	0.67	0.88

13. In a trial of two types of rain gauges, 65 of type A and 65 of type B were distributed at random over a certain region. In a given period, 14 storms occurred, and the average amounts of rain found in the two types of gauges were as shown below (from Brownlee, 1960). Test the hypothesis that the two types of gauges are giving similar results, using the sign test. Use $\alpha = 0.05$.

Storm:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Type A:	1.38	9.69	0.39	1.42	0.54	5.94	0.59	2.63	2.44	0.56	0.69	0.71	0.95	0.5
Type B:	1.42	10.37	0.39	1.46	0.55	6.15	0.61	2.69	2.68	0.53	0.72	0.72	0.93	0.53

14. The following data show the weight (in lb) and systolic blood pressure of 10 randomly selected male adults. Do the data provide sufficient evidence to indicate that there is a positive association between weight and systolic blood pressure? Use $\alpha = 0.05$.

Weight (lb):	198	160	192	198	167	176	150	151	168	169
Systolic BP:	135	145	138	145	136	153	133	139	137	157

15. A random sample of 12 students was selected from the freshman class at a liberal arts college. For each student, final scores in a first course in both calculus and physics are recorded and appear as given below. Do these data provide sufficient evidence to indicate that there is an association between calculus and physics scores? Use $\alpha = 0.01$.

Calculus:	95	96	94	94	87	88	89	90	95	93	92	93
Physics:	91	91	95	90	82	95	85	84	83	87	87	92

Chapter 15

SIMPLE LINEAR REGRESSION ANALYSIS

The focus of this chapter is the development of some procedures employed in simple linear regression analysis.

Topics Covered

- Basic concepts of regression analysis
- Fitting a straightline by least squares
- Unbiased estimation of error variance σ^2
- Tests and confidence intervals for the regression coefficients β_0, β_1 of the simple linear regression model
- Determination of confidence intervals for $E(Y|X)$
- Determination of a prediction interval for a future observation Y
- Inference about the correlation coefficient ρ
- Residual analysis

Learning Outcomes

After studying this chapter, the reader will be able to

- Fit a simple linear regression model to a given set of data and perform a residual analysis to check the validity of the model under consideration.
- Estimate the regression coefficients using the method of least squares and carry out hypothesis testing to test whether the first-order regression model is an appropriate fit to the given data.
- Estimate the expected response, predict future observation values, and find their confidence intervals using the given confidence coefficients.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

- Make inferences about the correlation coefficient between the response variable and the predictor variables.
- Use statistical packages MINITAB,R, and JMP to perform regression analysis.

15.1 INTRODUCTION

In this chapter, and the next, we deal with aspects of mathematical model building for the purpose of either describing a natural phenomenon based on some observable variables or predicting the value of a variable with the help of the observed values of one or several other variables. The reader may be familiar with the hypothesis (Boyle's law for gases) stating that (pressure) × (volume) = constant under a given temperature. This hypothesis is formulated after taking several observations on pressure and volume and noting that Boyle's law holds. Thus, a natural phenomenon is described by the following mathematical equation, called the *mathematical model*:

$$P \times V = C|T$$

where P, V, T denote the pressure, volume, and temperature, respectively, and C is a constant; the notation $C|T$ is read as "constant for a given temperature." Such a model is true until further evidence becomes available to prove that it is not true. In such cases, a new hypothesis must be formulated and further research is required to find a true model that then becomes the new law.

In this and the next chapter, we are not trying to create models that become laws to describe natural phenomena; rather, we are setting up models to help clearly describe a natural phenomenon or process and to predict a variable by observing other variables. In the model $P \times V = C|T$ given above, only two variables, P and V, are involved for a given temperature T.

Quite often, however, we are confronted with statistical problems where we must deal with three, four, or more variables. For example, consider the problem of estimating the weight of a male from his height. Suppose that n males are selected and their heights, X (feet), and weights, Y (pounds), are observed. Further, suppose that these observations indicate a relationship of the form $Y = 10 + 25X$. If this is the true relationship, then any other person's height and weight should satisfy the same model equation, $Y = 10 + 25X$. For example, if a person's height is 6 ft, then following the model, his weight should be 160 lb.

Suppose, however, that his actual weight is 180 lb, or a 20 lb difference between the predicted and the observed value. If a second person is weighed and measured and if his height and weight are checked with the model, then perhaps the difference between the predicted and observed values of the weight may differ from 20 lb. The reason for this difference between the predicted and observed values, as well as the deviations between the differences, may be attributable to numerous factors. For example, our sample of n persons may not be a representative sample of the population for which we want to create a prediction model. This discrepancy can be corrected by a proper random sampling procedure. Yet we may find differences between the predicted and observed values, as well as the deviations between the differences.

Many other factors may contribute to the weights, such as the amount of food consumed, genetic factors, amount of daily exercise, and climate or environmental factors. Suppose that all the recognizable factors that may have some effect on the weight are isolated and expressed as $X_1, X_2, \ldots, X_k$. If these are the only factors, and if there exists a function $f(X_1, X_2, \ldots, X_k)$ giving the weight Y, the problem simplifies to one of determining f based on a properly selected random sample of measurements on $X_1, X_2, \ldots, X_k$ and Y.

Suppose, by some technique, the form of f is constructed. Then the measurements $X_1, X_2, \ldots, X_k$ should predict the weight Y as $f(X_1, X_2, \ldots, X_k)$. Often this predicted value may not again agree with the observed weight. This means that either the construction of the model $f(X_1, X_2, \ldots, X_k)$ is not properly done or there is a number of other unknown factors that contribute to a person's weight. The sum total contribution from all unknown factors is called the "random part" in the model; it is usually denoted by ε. Thus the model becomes

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon \tag{15.1.1}$$

where f and ε are unknown, and $X_1, X_2, \ldots, X_k$ and Y are observable. The two basic problems are (i) defining f and (ii) constructing f. In statistical terminology, the model in (15.1.1) is the *regression* model with f the regression function, the variable Y the *response* or *dependent* variable, and the variables $X_1, X_2, \ldots, X_k$ the *predictor* or *independent* variables. In this chapter, we consider a simple aspect of constructing the regression model using the statistical packages, MINITAB, R, and JMP, when there is only one predictor variable involved.

15.2 FITTING THE SIMPLE LINEAR REGRESSION MODEL

15.2.1 Simple Linear Regression Model

Consider the regression model (15.1.1) when there is only one independent variable and the regression function is linear. Thus the model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{15.2.1}$$

From this model, we have

1. Y_i is the value of the response variable in the ith trial, or experiment.
2. β_0 and β_1 are the unknown parameters, usually referred to as regression coefficients. Here β_0 is the y intercept (value of Y when $X = 0$) and β_1 is the slope of the line; that is, the rate of change in Y as X changes.
3. X_i is the value of the predictor (or independent) variable selected in the ith trial, at which we observe Y_i.
4. The ε_i's are random variables, and $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$, with ε_i, ε_j (for all $i, j; i \neq j$), $i, j = 1, 2, \ldots, n$ uncorrelated so that, $E(\varepsilon_i \varepsilon_j) = 0$.

Note that Y_i is a value of an observable random variable, X_i is a nonrandom preselected value of X used to generate Y_i, and ε_i is an unobservable random variable. Also β_0 and β_1, the so-called regression parameters, are unknown, so that estimates of β_0 and β_1 are desired. The model (15.2.1) is called the *simple linear regression* model because it is linear in regard to the parameters β_0 and β_1. Furthermore, since the model is also linear in the predictor variable X, this regression model is called a *first-order linear model*. Now, in view of condition 4 on the random variable ε_i, the model (15.2.1) may alternatively be expressed as

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i, \quad i = 1, 2, \ldots, n \tag{15.2.2}$$

and, in general, we often refer to the model (15.2.2) as

$$E(Y) = \beta_0 + \beta_1 X \tag{15.2.3}$$

Letting $E(Y|X) = \eta$, the model (15.2.3) may be written as

$$\eta = \beta_0 + \beta_1 X \tag{15.2.3a}$$

This model relates the mean of the dependent variable Y linearly to the independent variable X. Now, to achieve the goal of constructing the regression model, we need to estimate the unknown parameters, β_0 and β_1. The process of estimating these parameters and evaluating them based on the information contained in a random sample of n ordered pairs (x_i, y_i), $i = 1, \ldots, n$, of observations is called *regression analysis*.

Before we start estimating the unknown parameters β_0 and β_1 of the regression model (15.2.1), we must evaluate whether, for the given set of observations (x_i, y_i), the model is viable. One way to evaluate whether the model is viable is to draw a scatter plot for the given set of observations (x_i, y_i), and "eyeball" a straightline through the data points to see if a first-order model gives a reasonable fit. There may be no apparent relationship, indicating the model is not viable, or there may be an apparent relationship, indicating that the model is viable.

Example 15.2.1 (Gas volume versus temperatures) *A gas can be kept at a constant pressure by placing it in a cylinder with a free moving piston. The volume of gas in such a cylinder increases with temperature. An experiment can be performed in which the volume of the gas in the cylinder is measured at various preselected settings of temperature. A record and plot of such data appear in Figure 15.2.1.*

Temperature °C: X	0	20	40	60	80	100
Volume cm^3: Y	50.0	53.7	57.3	61.0	64.8	68.3

It is obvious that the plotted points fall along a straightline. Students of engineering will recognize this relationship as Charles's law. In this example, the simple linear regression model is clearly appropriate. However, in problems where the plotted points do not fall nearly along a straight line, the problem of analysis becomes more complicated. The difficulty of drawing a line through a scattering of points, $(X_i, Y_i), i = 1, \ldots, n$, in Figure 15.2.1, is that repeating this experiment on the same set of points will likely result in quite different results due to the randomness of the Y_i's. What is clearly needed is some objective way to fit a straight line to observed data, or to estimate the unknown

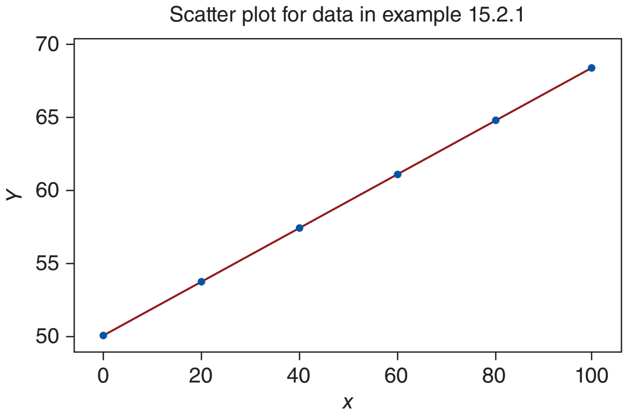


Figure 15.2.1 Experimental verification of Charles's law.

parameters β_0 and β_1. The most widely used method for estimating the parameters β_0 and β_1, is the method of *least squares*, which we discuss in Section 15.2.2.

So far we have not considered any particular form of the probability distribution of the random error ε. It should also be noted that regardless of any form of the probability distribution of the random error ε, the least-squares method results in point estimators of β_0 and β_1 that are unbiased and have minimum variance among all unbiased linear estimators. This result is known as the *Gauss–Markov theorem.* However, any further inference on β_0 and β_1, such as determining confidence intervals and/or testing hypotheses, assumes that the random errors ε_i's are normally distributed as $N(0, \sigma^2)$. Under this assumption that ε_i's are normally distributed, the model (15.2.1) is written as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{15.2.4}$$

where we assume that:

1. Y_i is the value of the response variable in the ith trial, and Y_i's are independent random variables.
2. β_0 and β_1 are the unknown parameters, where β_0 is the y-intercept (value of Y when X is zero) and β_1 is the slope of the line, that is, the rate of change in Y as X changes. The parameters β_0 and β_1 are often called the regression coefficients.
3. X_i is the value of the predictor variable in the ith trial.
4. ε_i's are independent $N(0, \sigma^2)$ random variables.

Note that under the assumption that ε_i's are normally distributed, the condition that the ε_i's are uncorrelated in (15.2.1) turns into the assumption of independence, since two uncorrelated normal random variables are independent (Chapter 5). The model (15.2.4) when assumptions 1–4 hold is also sometimes called the *normal error regression model.*

Furthermore, using model (15.2.4) and some of the characteristics of the normal distribution, we can easily rewrite (15.2.4) as

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2), \quad Y_i\text{'s independent} \tag{15.2.5}$$

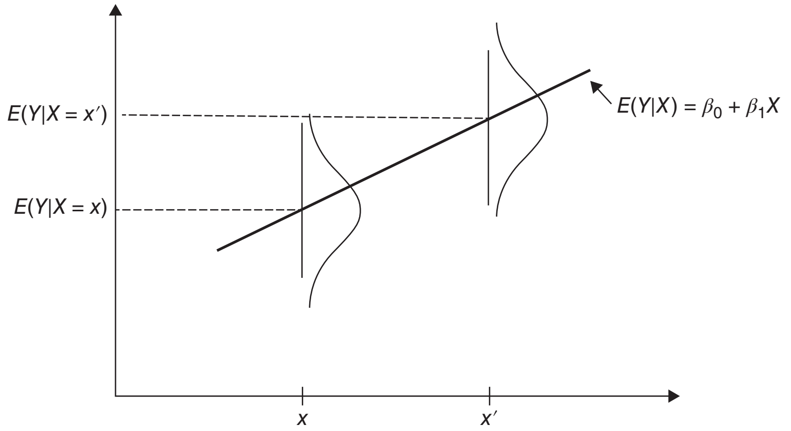


Figure 15.2.2 Distributions of Y at X having values x and x' have exactly the same form, but differing in location.

Note that $Var(Y_i)$ has the same value σ^2 for all Y_i. In other words, for different X_i, the probability distributions of Y_i are the same except for the location parameter. Figure 15.2.2 illustrates this concept.

15.2.2 Fitting a Straight Line by Least Squares

Suppose that the true relation between a response Y (e.g., hardness of steel) and a controlled variable X (e.g., X is the amount of carbon molecules inserted into the crystal structure of steel as it cools) is linear so that it can be modeled by the straightline model in (15.2.1). Let us agree to select several different values of X before the experiment, say $(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)$, and record the observation y_i of Y_i at each of the values of X (some values of X may be repeated). The choice of $(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)$ is referred to as the *design of the experiment.* We assume that $E(Y|X) = \eta$ and that the variance σ^2 of Y is independent of the values of X.

Now the n pairs of measurements $(x_i, y_i), i = 1, \ldots, n$, may be plotted as points in the $(x,\ y)$-plane, thus producing a scatter diagram (see Figure 15.2.3). We want to estimate β_0, β_1, σ^2, and we will use the information contained in the n sample points to estimate these parameters.

To find good (in some sense) estimators of the regression parameters β_0 and β_1, we use the method of least squares. The least-squares method produces the smallest possible sum of squared errors, the squared deviations of the observed Y_i from the estimate of their true means $E(Y_i) = \beta_0 + \beta_1 X_i$. In other words, the least-squares method proceeds by minimizing, over choices of β_0 and β_1, the quantity

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{15.2.6}$$

From (15.2.4) and (15.2.5) we have that the probability density of Y, given $X = x$, is

$$f(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2\right) \tag{15.2.7}$$

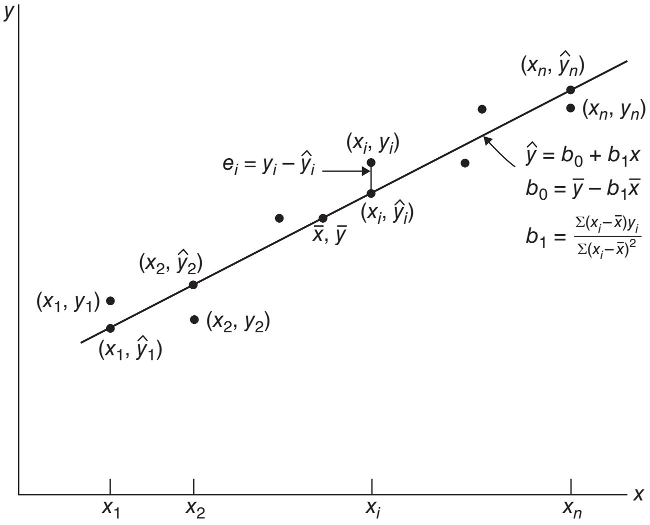


Figure 15.2.3 Scatter plot of n pairs of measurements $(x_i, y_i), i = 1, \ldots, n$, on hardness of certain steels.

Further, from our earlier discussion in this section, we have assumed that the random variables $Y_1, Y_2, \ldots, Y_n$ are statistically independent so that the likelihood function is the joint probability density function based on the sample n pairs $(x_1,\ y_1)$, $(x_2,\ y_2)$, $\ldots$, $(x_n,\ y_n)$ and is given by

$$\begin{aligned}\textstyle\prod_{i=1}^{n} f(y_i|x_i) &= \textstyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right) \\ &= \left\{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)\right\}\end{aligned} \tag{15.2.8}$$

As discussed in Chapter 8, a widely used method for estimating parameters is to use those values that maximize the likelihood function. If we apply this method of maximum likelihood to finding estimators of β_0, β_1, and σ^2, we must maximize the quantity within the brackets{ } in (15.2.8). Note that maximizing the quantity within the brackets { } with respect to β_0 and β_1 involves the problem of minimizing with respect to β_0 and β_1 the sum of squares function Q of (15.2.6), namely

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{15.2.9}$$

located in the exponential part of (15.2.8). This is because of the presence of the minus sign in the exponential part of (15.2.8). Hence, under the assumption of normality of the errors ε_i (see condition 4 of (15.2.4)), we find that the maximum likelihood estimators and the least-squares estimators of β_0 and β_1 are one and the same.

Note that we could have used the principle of least squares to find estimators *without invoking the condition of normality.* Interestingly, it can be proved that under the assumptions 1–3 of (15.2.4), minimum-variance, unbiased estimators of parameters that are linear combinations of independent random variables are provided by the method of least squares. As mentioned earlier, this is a theorem due to the famous mathematician Karl Gauss.

Further, it should be noted that in dealing with the mathematical models (15.2.3a) and (15.2.4), the variable X is assumed controlled, or measured, without error, and all the random variability is attributed to the observation Y at each fixed value of X. In practice, controlling or measuring X *exactly* is often impossible. However, it is assumed that the values of X are known much more precisely than those of the response, Y, and indeed the analysis proceeds by assuming the independent variable, X, can be fixed without error.

To find estimators of β_0 and β_1, that is, the values of (β_0, β_1) that minimize $Q(\beta_0, \beta_1)$ in (15.2.9), say b_0 and b_1, we proceed by taking the partial derivatives with respect to β_0 and β_1 and setting these derivatives equal to zero. We then have

$$\left.\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0}\right|_{b_0, b_1} = 0, \qquad \left.\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1}\right|_{b_0, b_1} = 0 \tag{15.2.10}$$

This in turn produces the two equations

$$\begin{aligned} -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) &= 0 \\ -2\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) &= 0 \end{aligned} \tag{15.2.11}$$

Dividing each of these equations by -2 and performing the summations, remembering that $\sum_{i=1}^{n} Y_i = n\bar{Y}$ and $\sum_{i=1}^{n} X_i = n\bar{X}$, we obtain two equations in b_0 and b_1 given by

$$n(b_0) + n\bar{X}(b_1) = n\bar{Y} \quad \text{and} \quad n\bar{X}(b_0) + \sum X_i^2(b_1) = \sum X_i Y_i \tag{15.2.12}$$

Equation (15.2.12) are usually called the *normal equations* for estimating β_0 and β_1. They are said to be *in standard form* in that b_0 and b_1 are mentioned only on the left-hand sides of Equation (15.2.12).

Thus, by solving the two Equations in (15.2.12) with respect to (b_0, b_1), we obtain the solution applicable to $(X_1, Y_1), \ldots, (X_n, Y_n)$, namely to a set of $(X_1, \ldots, X_n)$ to be chosen at which we will observe $(Y_1, \ldots, Y_n)$, respectively. Solving Equation (15.2.12) for b_0 and b_1, we obtain

$$\begin{aligned} b_0 &= \bar{Y} - b_1\bar{X}, \text{where} \\ b_1 &= \tfrac{\sum_{i=1}^{n} X_i\ Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \tfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \tfrac{S_{XY}}{S_{XX}} \end{aligned} \tag{15.2.13}$$

Here S_{XY} is called the *corrected sums of cross products*, and S_{XX} and S_{YY} are called the *corrected sum of squares*. The reader should verify that

$$S_{XX} = \sum X_i^2 - n\bar{X}^2, \quad S_{XY} = \sum X_i Y_i - n\bar{X}\ \bar{Y} \tag{15.2.14}$$

and

$$S_{YY} = \sum Y_i^2 - n\bar{Y}^2 \tag{15.2.15}$$

Given the estimators b_0 and b_1, the *fitted model* is

$$\hat{Y} = b_0 + b_1 X = (\bar{Y} - b_1\bar{X}) + b_1 X = \bar{Y} + b_1(X - \bar{X}) \tag{15.2.16}$$

The estimation procedure above is often called *regressing Y upon X*. Equation (15.2.16) is often called the fitted regression line or regression equation. We illustrate this method in

Example 15.2.2 below. But, at this point, we note that as long as the determinant of the coefficients in Equation (15.2.12), is not zero, that is, as long as the determinant

$$\begin{vmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{i=1}^{n} X_i^2 \end{vmatrix} = n\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \neq 0$$

Equation (15.2.12) have a unique solution with respect to (b_0, b_1). Now the determinant has the value $n\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$, that is, $n\sum(X_i - \bar{X})^2$, which will not be zero unless all X_i are equal. As stated at the outset, equality of all X_i is ruled out.

Example 15.2.2 (Steel hardness versus carbon content) *One of the factors that determines the hardness of steel is the carbon content. The carbon molecules insert themselves in the crystal structure of steel as it cools. Molecules of carbon and steel are much harder to break than steel alone. The data below give the hardness of steel in units of 1000 psi and the percentage of carbon contents:*

Steel hardness: Y	76	79	78	86	77	80	86	79	75	82
Carbon content: X	0.21	0.23	0.24	0.31	0.23	0.28	0.33	0.24	0.22	0.26

(a) Construct a scatter plot for these data and draw a straight-line to examine if the simple linear regression model seems to be an adequate model for these data.

(b) Find the fitted regression line for these data.

Solution: (a) The MINITAB printout of the scatter plot for this example is shown in Figure 15.2.4. The scatter plot in Figure 15.2.4 clearly indicates that a simple linear regression model seems to be an adequate model for the given data.

(b) We now proceed to find the regression line for these data.

From the data we have $n = 10$ and

$$\begin{aligned} \bar{X} &= \tfrac{1}{10}\sum_{i=1}^{n} X_i = \tfrac{2.55}{10} = 0.2550, \quad \bar{Y} = \tfrac{1}{10}\sum_{i=1}^{n} Y_i = \tfrac{798}{10} = 79.80 \\ S_{XX} &= \sum X_i^2 - n\bar{X}^2 = 0.6645 - 10(0.2550)^2 = 0.01425 \\ S_{XY} &= \sum X_i Y_i - n\bar{X}\bar{Y} = 204.78 - 10(0.2550) \times (79.80) = 1.29 \end{aligned}$$

From (15.2.13) we have

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{1.29}{0.01425} = 90.526, \quad b_0 = \bar{Y} - b_1\bar{X} = 79.8 - (90.526)(0.255) = 56.716$$

Thus the fitted regression line is

$$\hat{Y} = 56.716 + 90.526X$$

Note that in the example $\hat{Y}$ given in (15.2.16) is an estimator of the regression line $E(Y|X) = \beta_0 + \beta_1 X$, and also gives a point estimator of an observation Y, taken at X. If the model $E(Y|X) = \beta_0 + \beta_1 X$ is written in the equivalent alternative form $E(Y|X) = \delta + \beta_1 X'$, where $X' = (X - \bar{X})$ and $\delta = \beta_0 + \beta_1\bar{X}$, the normal equations for d and b_1, the estimators of δ and β_1 are given by

$$(n)d + (0)b_1 = \sum_{i=1}^{n} Y_i \quad \text{and} \quad (0)d + \left(\sum X'^2\right) b_1 = \sum_{i=1}^{n} X_i' \, Y_i \qquad (15.2.17)$$

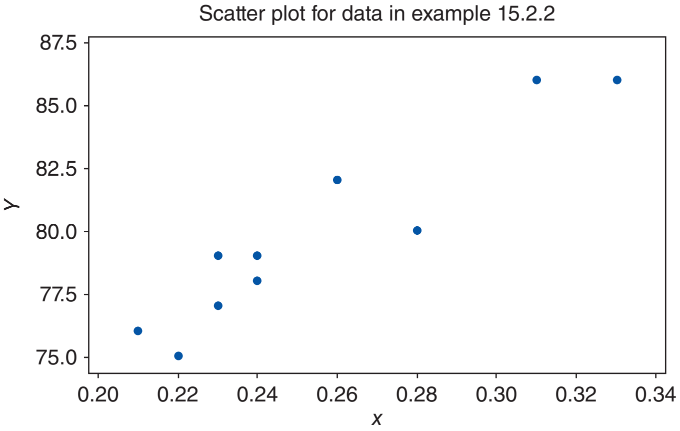


Figure 15.2.4 Scatter plot for the data in Example 15.2.2.

Solution of these normal equations results in

$$d = \frac{\sum Y_i}{n} = \bar{Y}, \quad b_1 = \frac{\sum_{i=1}^{n} X_i' Y_i}{\sum_{i=1}^{n} X_i'^2} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum X_i'^2} = \frac{\sum (X_i - \bar{X})\,(Y_i - \bar{Y})}{\sum\,(X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} \tag{15.2.18}$$

since $\sum_{i=1}^{n} X_i' = \sum_{i=1}^{n} (X_i - \bar{X}) = 0$, so that $\sum_{i=1}^{n} (X_i - \bar{X})\bar{Y} = 0$.

The fitted line then becomes $\hat{Y} = \bar{Y} + b_1 X'$, and the exchange of X' for $X - \bar{X}$ effectively shifts the origin to the average of the X_i's. The intercept of the straight-line in the new coordinate system is then simply the average $\bar{Y}$. The estimate of β_0 in the original model can be retrieved. On finding $(d, b_1) = (\bar{Y}, b_1)$ from (15.2.18), we could find an estimate of β_0 by simply evaluating $\bar{Y} - b_1\bar{X}$, which from (15.2.13) is the estimate b_0. Indeed, as we will see, $E(b_0) = E(\bar{Y} - b_1\bar{X}) = \beta_0$ and $E(b_1) = \beta_1$; that is, b_0 and b_1 are unbiased estimators of β_0 and β, respectively. Rewriting the original model in this modified form simplifies computations. The quantity $X' = (X - \bar{X})$ is sometimes called the *orthogonal polynomial of degree one.*

Graphically, we can represent all possible values of X and the corresponding values of $\hat{Y}$, that is, all possible pairs $(X, \hat{Y})$, as a straight-line in the (x, y)-plane whose equation is given by (15.2.16). As mentioned before, this line is called the regression line of Y on X with intercept b_0 and slope b_1. From (15.2.16) we see that the regression line passes through the point $(\bar{X}, \bar{Y})$, the *center of gravity* of the scatter diagram of the n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$.

15.2.3 Sampling Distribution of the Estimators of Regression Coefficients

We assume in this section that (15.2.4) holds so that $Y_1, Y_2, \ldots, Y_n$ are independent random variables having normal distributions $N(\beta_0 + \beta_1 X_1, \sigma^2)$, $N(\beta_0 + \beta_1 X_2, \sigma^2)$, $\ldots$, $N(\beta_0 + \beta_1 X_n, \sigma^2)$ where $X_1, \ldots, X_n$ are constants (not random variables). We then have, using (15.2.13), that the statistic b_1 is a linear function of the observations $Y_1, \ldots, Y_n$ with coefficients

$$\frac{X_1 - \bar{X}}{\sum(X_i - \bar{X})^2}, \ldots, \frac{X_n - \bar{X}}{\sum(X_i - \bar{X})^2} \tag{15.2.19}$$

That is, b_1 is a statistic of the form $c_1Y_1 + c_2Y_2 + \cdots + c_nY_n$, where the coefficients c_i are n constants given by $c_i = (X_i - \bar{X})/\left[\sum(X_i - \bar{X})^2\right] = (X_i - \bar{X})/S_{XX}$. Thus, we have the important result that b_1 is a random variable having a normal distribution, that is,

$$(b_1|X_1, \ldots, X_n) \sim N\left(\beta_1; \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \quad \text{or} \quad (b_1|X_1, \ldots, X_n) \sim N\left(\beta_1; \frac{\sigma^2}{S_{XX}}\right) \tag{15.2.20}$$

To verify that the expected value of b_1 is β_1, we first note that

$$\sum_{i=1}^{n} c_i = 0, \quad \sum_{i=1}^{n} c_iX_i = 1$$

This result holds since

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{S_{XX}} = \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

and

$$\sum_{i=1}^{n} c_iX_i = \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X})X_i = \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{S_{XX}}{S_{XX}} = 1$$

Hence, we have that

$$\begin{aligned} E(b_1) &= E(c_1Y_1 + \cdots + c_n\,Y_n) = c_1E(Y_1) + \cdots + c_nE(Y_n) \\ &= c_1(\beta_0 + \beta_1X_1) + \cdots + c_n(\beta_0 + \beta_1X_n) \\ &= \beta_0(c_1 + \cdots + c_n) + \beta_1(c_1X_1 + \cdots + c_nX_n) = \beta_1 \end{aligned} \tag{15.2.21}$$

Then, to see that the variance of b_1 is $\sigma^2/\sum(X_i - \bar{X})^2$, we have

$$V(b_1) = \sigma^2_{b_1} = c_1^2\sigma^2_{Y_1} + \cdots + c_n^2\sigma^2_{Y_n} = (c_1^2 + \cdots + c_n^2)\sigma^2$$

Now

$$\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n}\left(\frac{(X_i - \bar{X})}{S_{XX}}\right)^2 = \frac{1}{(S_{XX})^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{S_{XX}}{(S_{XX})^2} = \frac{1}{S_{XX}}$$

Hence we can state that

$$E(b_1) = \beta_1 \quad \text{and} \quad V(b_1) = \tfrac{1}{S_{XX}}\sigma^2 \tag{15.2.22}$$

Further, we have

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1X_i) = \beta_0 + \beta_1\bar{X}$$

Also from (15.2.13) we have $b_0 = \bar{Y} - b_1\bar{X}$ so that

$$E(b_0) = E(\bar{Y} - b_1\bar{X}) = E(\bar{Y}) - E(b_1\bar{X}) = E(\bar{Y}) - \bar{X}E(b_1) = \beta_0 + \beta_1\bar{X} - \bar{X}\beta_1 = \beta_0 \tag{15.2.23}$$

so that b_0 is an unbiased estimator of β_0. Now from (15.2.13), it follows that the statistic b_0, a linear function of the observations $Y_1, \ldots, Y_n$ with coefficients

$$\left(\frac{1}{n} - \frac{(X_1 - \bar{X})\bar{X}}{\sum (X_i - \bar{X})^2}\right), \cdots, \left(\frac{1}{n} - \frac{(X_n - \bar{X})\bar{X}}{\sum (X_i - \bar{X})^2}\right) \tag{15.2.23a}$$

But because $Y_1, \ldots, Y_n$ are independent and normally distributed, we can easily show, by using the same argument for determining the sampling distribution of b_1, that

$$V(b_0) = \sigma^2_{b_0} = \left(\tfrac{1}{n} + \tfrac{\bar{X}^2}{S_{XX}}\right)\sigma^2 \tag{15.2.24}$$

Hence,

$$(b_0|X_1, \ldots, X_n) \sim N\left(\beta_0; \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)\sigma^2\right) \tag{15.2.25}$$

Now lastly, from the earlier results of this section, we have that

$$E(\hat{Y}) = E(b_0 + b_1X) = E(b_0) + E(b_1)X = \beta_0 + \beta_1X$$

We can then write that

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$

as an unbiased estimator of $\eta = E(Y|X) = \beta_0 + \beta_1X$.

Using the result (not proved here) that the covariance between $\bar{Y}$ and b_1 is zero, we have

$$V(\hat{Y}) = V(\bar{Y}) + (X - \bar{X})^2V(b_1) = \left(\tfrac{1}{n} + \tfrac{(X-\bar{X})^2}{S_{XX}}\right)\sigma^2 \tag{15.2.26}$$

PRACTICE PROBLEMS FOR SECTION 15.2

1. A team of physicians claims that a person's excessive weight adversely affects his/her plasma glucose level. The data they obtained on 10 "overweight" persons are given below:

Weight (lb), X:	181	180	189	188	229	192	223	231	212	225
Plasma glucose (mg/dL), Y:	206	162	181	199	214	146	210	165	177	183

(a) Construct a scatter plot for these data. Does the scatter plot indicate that a least-squares line would provide a good fit?

(b) Fit a regression line to these data for predicting the plasma glucose level.

(c) Using (b), record the least-squares estimates of the regression coefficients.

2. A chemist is interested in investigating the relationship between the reaction time and the yield of a chemical. He conducted 11 experiments with varying reaction time and measured the yield of the chemical. The data obtained are given below:

Reaction time, X:	33	43	43	32	30	43	30	38	40	40	32
Chemical yield, Y:	84	78	84	79	75	75	89	86	87	88	85

   (a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
   (b) Fit a regression line to these data for predicting the chemical yield.
   (c) Using (b), record the least-squares estimates of the regression coefficients.

3. A chemical engineer wants to investigate the relationship between the cooking time of paper pulp and the shear strength of the paper. She arranges to collect data for 12 batches of pulp that are cooked at the same temperature but for different periods of time. The data (coded) obtained is given below:

Cooking time, X:	14	11	10	13	12	13	12	10	14	15	12	13
Shear strength, Y:	125	70	97	86	111	121	89	130	99	95	125	102

   (a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
   (b) Fit a regression line to these data for predicting the shearing strength of the paper.
   (c) Using (b), record the least-squares estimates of the regression coefficients.

4. Recent studies show that high sound level (in decibels) makes humans prone to hypertension and heart attacks. For example, normal conversation level is 60 dB, for textile looms it is 105 dB, and for pneumatic chippers it is 115 dB. The following coded data give the noise level and the hypertension for people who work in noisy places:

Noise level, X:	28	33	21	35	29	26	22	30	34	27	31	34	29
Hypertension, Y:	73	68	69	88	80	74	74	69	89	68	76	87	73

   (a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
   (b) Fit a regression line to these data for predicting the hypertension level of a person.
   (c) Using (b), record the least-squares estimates of the regression coefficients.

5. The following data give the highest daytime temperature °F and amount of rainfall (in inches) on 10 randomly selected summer days at a tourist place in the northeastern United States:

Temperature, X:	91	90	96	94	87	90	93	81	94	91
Rain, Y:	0.8	0.4	0.5	0.7	0.3	0.4	0.6	0.1	0.4	0.7

(a) Construct a scatter plot for these data. Does the scatter plot indicate a straight-line would provide a good fit?
(b) Fit a regression line to these data for predicting the amount of rainfall.
(c) Using (b), record the least-squares estimates of the regression coefficient.

6. In the paper manufacturing process, too much moisture left in the paper causes streaks that renders the paper unusable. Streaks can be avoided, for example, by slowing down the machine, over drying the paper, calibrating the dryer head, and so on. The following coded data give the amount of moisture in the paper and the number of streaks per 100 linear feet:

Moisture, X:	8	11	14	9	10	13	12	7
Number of streaks, Y:	5	14	22	16	17	8	19	10

(a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
(b) Fit a regression line to these data for predicting the number of streaks.
(c) Using (b), record the least-squares estimates of the regression coefficients.

7. The following data give the final scores for 12 randomly selected students in courses on probability and operations research (OR):

Probability, X:	91	77	82	78	73	88	96	75	92	95	78	82
OR, Y:	86	75	86	76	75	89	87	91	83	90	84	94

(a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
(b) Fit a regression line to these data for predicting the scores in the OR course.
(c) Using (b), record the least-squares estimates of the regression coefficients.

8. In agriculture, it is important for obtaining a higher yield of certain crops that, at the time of planting, a proper distance be kept between plants. An agronomist conducted an experiment to investigate this for cotton crops. He divided a piece of land into small squares and sowed cotton seeds in each square. The following data give the distance X (in inches) between cotton plants and the yield Y (in pounds per square inch):

Yield, Y:	73	88	75	88	101	116	94	72
Distance, X:	12	15	13	14	18	19	15	12

(a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
(b) Fit a regression line to these data for predicting the yield of cotton.
(c) Using (b), record the least-squares estimates of the regression coefficients.

9. Copper wire is the most widely used conductor since it has high conductivity and good mechanical properties. A wide range of cable applications require high tensile strength for copper wires. Mixing beryllium with copper increases the tensile

strength of copper wires. The following (coded) data give the percentage of beryllium mixed with copper (X) and tensile strength of copper wire (Y):

Tensile strength, Y:	50	64	67	59	74	63	56	61	69
Beryllium, X:	5	7	8	6	9	6	5	7	8

(a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight-line would provide a good fit?
(b) Fit a regression line to these data for predicting the tensile strength of copper wire.
(c) Using (b), record the least-squares estimates of the regression coefficients.

10. The purity (%) of oxygen produced by a fractional distillation process is believed to be related to percentage of hydrocarbons (%) in the main condenser of the processing unit. The data obtained on 20 samples are given below (from *Introduction to Linear Regression Analysis* by Montgomery et al. (2006), used with permission):

Purity, Y:	86.91	89.85	90.28	86.34	92.58	87.33	86.29	91.86	95.61	89.86
Hydrocarbon, X:	1.02	1.11	1.43	1.11	1.01	0.95	1.11	0.87	1.43	1.02
Purity, Y:	96.73	99.42	98.66	96.07	93.65	87.31	95.00	96.85	85.20	90.56
Hydrocarbon, X:	1.46	1.55	1.55	1.55	1.40	1.15	1.01	0.99	0.95	0.98

(a) Construct a scatter plot for these data. Does the scatter plot indicate that a straight line would provide a good fit?
(b) Fit a regression line to these data for predicting the purity of oxygen.
(c) Using (b), record the least-squares estimates of the regression coefficients.

11. The weight (X) and total cholesterol level (Y) of 20 randomly selected females in the age group 30 and 40 are given below. Assume that the simple linear regression model is appropriate for these data.

Weight, X:	135	140	128	143	150	155	147	146	137	144
Cholesterol, Y:	154	152	149	140	165	169	154	152	139	133
Weight, X:	126	134	143	149	146	141	138	152	151	145
Cholesterol, Y:	132	136	139	150	154	149	160	166	164	158

(a) Construct a scatter diagram for these data.
(b) Fit a simple linear regression model to these data.
(c) Using (b), record the least-squares estimates of the regression coefficients.

12. A study was made on the effect of temperature (X) on the yield (Y) of a chemical process. The following data (coded) were collected (from Draper and Smith, 1981, used with permission):

X:	−5	−4	−3	−2	−1	0	1	2	3	4	5
Y:	1	5	4	7	10	8	9	13	14	13	18

(a) Construct a scatter diagram for these data.
(b) Fit a simple linear regression model to these data considering X an independent variable and Y a response variable. Assume that the simple linear regression model is appropriate for these data.
(c) Using (b) record the least-squares estimates of the regression coefficients.

15.3 UNBIASED ESTIMATOR OF σ^2

In Section 15.2, we developed the sampling distributions of the statistics b_0 and b_1 that depend on unknown σ^2. We now turn to the problem of finding an unbiased estimator for σ^2. We denote the value on the regression line corresponding to X_i as $\hat{Y}_i$, that is, $\hat{Y}_i = b_0 + b_1 X_i = \bar{Y} + b_1(X_i - \bar{X})$. Suppose that we now take the differences, often called residuals, between the observed value of Y at X_i (i.e., Y_i) and the fitted value of Y at X_i (i.e., $\hat{Y}_i$) for $i = 1, 2, 3, \ldots, n$. Denoting these differences by $e_i = Y_i - \hat{Y}_i$, we may square and then sum the e_i, denoting the result by SSE, the *sum of squares of errors*. We have

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_i)^2 \tag{15.3.1}$$

The quantity $e_i = Y_i - \hat{Y}_i$ is often called the *residual* of Y_i, and the sum of squares in (15.3.1) is often called the *residual sum of squares* or *error sum of squares*. Since b_0 and b_1 are the least-squares estimators of β_0 and β_1, SSE is thus the minimum of $Q(\beta_0, \beta_1)$, where $Q(\beta_0, \beta_1)$ is given in (15.2.9). Now, taking the expected value of SSE, we have

$$\begin{aligned}
E(SSE) &= E\left(\sum (Y_i - \hat{Y}_i)^2\right) \\
&= E\left(\sum (Y_i - b_0 - b_1 X_i)^2\right) \\
&= E\left(\sum [(Y_i - \bar{Y}) - b_1(X_i - \bar{X})]^2\right) \\
&= E\left(\sum (Y_i - \bar{Y})^2 + b_1^2 \sum (X_i - \bar{X})^2 - 2b_1 \sum (X_i - \bar{X})(Y_i - \bar{Y})\right)
\end{aligned}$$

Recalling from (15.2.13) that $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = b_1 \sum_{i=1}^{n} (X_i - \bar{X})^2$ we have

$$\begin{aligned}
E(SSE) &= E\left(\sum (Y_i^2 - n\bar{Y}^2) - b_1^2 \sum (X_i - \bar{X})^2\right) \\
&= \sum E(Y_i^2) - nE(\bar{Y}^2) - \sum (X_i - \bar{X})^2 E(b_1^2)
\end{aligned}$$

Now, recalling that $E(W^2) = \mathrm{Var}(W) + [E(W)]^2$, we can rewrite the expression above as

$$\begin{aligned}
E(SSE) &= \sum [\sigma^2 + (\beta_0 + \beta_1 X_i)^2] - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{X})^2\right) \\
&\quad - \sum (X_i - \bar{X})^2 \left(\frac{\sigma^2}{\sum (X_i - \bar{X})^2} + \beta_1^2\right) \\
&= (n-2)\sigma^2 + \left\{\sum (\beta_0 + \beta_1 X_i)^2 - n(\beta_0 + \beta_1 \bar{X})^2 - \beta_1^2 \sum (X_i - \bar{X})^2\right\}
\end{aligned} \tag{15.3.2}$$

It is easily shown that the quantity in the braces { } in (15.3.2) is zero. Thus, we have the result

$$E(SSE) = (n-2)\sigma^2 \tag{15.3.3}$$

We emphasize again that (15.3.3) is based only on the assumptions 1–3 of (15.2.4). From (15.3.3) we have that an unbiased estimator of variance is given by

$$\boxed{\hat{\sigma}^2 = SSE/(n-2) \qquad (15.3.4)}$$

which we usually denote by S^2 or MSE. We call MSE the *error mean square* or *residual mean square*. The factor $(n-2)$ in the denominator is sometimes referred to as the degrees of freedom associated with the error sum of squares SSE.

Note that $SSE = \sum_{i=1}^{n} e_i^2$ and that

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \hat{Y}) = \sum_{i=1}^{n}(Y_i - (\bar{Y} + b_1(X_i - \bar{X}))) = \sum_{i=1}^{n}(Y_i - \bar{Y}) + b_1\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \tag{15.3.5}$$

Example 15.3.1 (Steel hardness versus carbon content) *Refer to Example 15.2.2 where we are given the data on hardness of steel and the carbon content. Recall that the carbon molecules insert themselves in the structure of steel as it cools. Find an unbiased estimate of σ^2.*

Solution: The regression line, as found in Example 15.2.2, is

$$\hat{Y} = 56.716 + 90.526X$$

and using the fitted values $\hat{Y}_i$ at X_i, the corresponding residuals are tabulated in Table 15.3.1.

From the residuals in Table 15.3.1, we obtain the value of $SSE = \sum(Y_i - \hat{Y}_i)^2$ as $SSE = 14.80$

Since the degrees of freedom associated with SSE are $10 - 2 = 8$, we obtain an unbiased estimate of σ^2 equal to

$$S^2 = 14.80/8 = 1.85$$

and an estimate of σ is $S = 1.36 = 1000(1.36) = 1360\,\text{psi}$ (the units of Y_i were in 1000 psi).

Table 15.3.1 Fitted values and the corresponding residuals for the data in Example 15.2.2.

Carbon contents, X	0.21	0.23	0.24	0.31	0.23	0.28	0.33	0.24	0.22	0.26
Steel hardness, Y	76	79	78	86	77	80	86	79	75	82
Fitted values, $\hat{Y}$	75.73	77.54	78.44	84.78	77.54	82.06	86.59	78.44	76.63	80.25
Residuals, $Y - \hat{Y}$	0.27	1.46	−0.44	1.22	−0.54	−2.06	−0.59	0.56	−1.63	1.75

Now using (15.3.4) the unbiased estimators of $\sigma^2_{b_0}$ and $\sigma^2_{b_1}$ are given by

$$\hat{\sigma}^2_{b_0} = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right) S^2 \tag{15.3.6}$$

and

$$\hat{\sigma}^2_{b_1} = \frac{S^2}{S_{XX}} \tag{15.3.7}$$

respectively.

An alternative form of SSE that is simpler for computational purposes is

$$SSE = S_{YY} - b_1 S_{XY} \tag{15.3.8}$$

where $S_{YY} = \sum Y_i^2 - n\bar{Y}^2 = \sum (Y_i - \bar{Y})^2$, and is sometimes referred to as the *corrected total sum of squares.* Also, $S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - n\bar{X}\,\bar{Y}$.

15.4 FURTHER INFERENCES CONCERNING REGRESSION COEFFICIENTS (β_0, β_1), $E(Y)$, AND Y

15.4.1 Confidence Interval for β_1 with Confidence Coefficient $(1 - \alpha)$

Since, from (15.2.20), we know that the probability distribution of the statistic b_1 under the normality of the Y_i's is such that

$$b_1 \sim N(\beta_1, \sigma^2/S_{XX})$$

Hence, by standardizing the random variable b_1, we find that

$$\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{XX}}} \sim N(0, 1) \tag{15.4.1}$$

Now, replacing σ in (15.4.1) by its estimator S, we obtain (see Chapter 7)

$$\frac{b_1 - \beta_1}{S/\sqrt{S_{XX}}} \sim t_{n-2} \tag{15.4.2}$$

This follows from the fact that under the assumptions 1–4 of (15.2.4), $(n-2)S^2/\sigma^2$ has the χ^2_{n-2} distribution. Using Equation (15.4.2), we can make the following probability statement:

$$P\left(-t_{n-2;\alpha/2} \le \frac{b_1 - \beta_1}{S/\sqrt{S_{XX}}} \le t_{n-2;\alpha/2}\right) = 1 - \alpha \tag{15.4.3}$$

After some algebraic manipulation, we find that

$$P\left(b_1 - t_{n-2;\alpha/2}\frac{S}{\sqrt{S_{XX}}} \le \beta_1 \le b_1 + t_{n-2;\alpha/2}\frac{S}{\sqrt{S_{XX}}}\right) = 1 - \alpha \tag{15.4.4}$$

or

$$\left(b_1 \pm t_{n-2;\alpha/2}\frac{S}{\sqrt{S_{XX}}}\right) \tag{15.4.4a}$$

is a confidence interval for β_1 with confidence coefficient $(1-\alpha)$.

Example 15.4.1 (Steel hardness versus carbon content) *Consider the data on hardness of steel in Example 15.2.2. Find a 95% confidence interval for β_1 by manual calculations (we discuss this using a statistical software package later in this chapter). Assume normality of the independent Y_i, where Y_i is observed at $X_i, i = 1, 2, \ldots, n$.*

Solution: From Examples 15.2.2 and 15.3.1, we have

$$b_1 = 90.526, \quad S_{XX} = 0.01425, \quad \text{and} \quad S = 1.36 = \hat{\sigma}$$

Since $n = 10$, $\alpha = 0.05$, from Table A.5 we have $t_{n-2;\alpha/2} = t_{8;.025} = 2.306$. We then have, from (15.4.4a), that the 95% confidence interval for β_1 is

$$\left(90.526 \pm 2.306\frac{1.36}{\sqrt{.01425}}\right) = (64.254, 116.798)$$

15.4.2 Confidence Interval for β_0 with Confidence Coefficient $(1-\alpha)$

Now, from (15.2.25), we know that under the normality of the Y_i's the probability distribution of the statistic b_0 is such that

$$b_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)\sigma^2\right) \tag{15.4.5}$$

Hence, by standardizing the random variable b_0, we have

$$\frac{b_0 - \beta_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim N(0,1) \tag{15.4.6}$$

Then, replacing σ in (15.4.6) by its estimator S, we find that

$$\frac{b_0 - \beta_0}{S\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim t_{n-2} \tag{15.4.7}$$

From Equation (15.4.7), we have that

$$P\left(-t_{n-2;\alpha/2} \leq \frac{(b_0-\beta_0)}{S\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}\right)}} \leq t_{n-2;\alpha/2}\right) = 1-\alpha \tag{15.4.8}$$

From (15.4.8) we find that

$$P\left(b_0 - t_{n-2;\alpha/2}S\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}\right)} \leq \beta_0 \leq b_0 + t_{n-2;\alpha/2}S\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}\right)}\right) = 1-\alpha \tag{15.4.9}$$

Hence the $100(1-\alpha)\%$ confidence interval for β_0 is

$$b_0 \pm t_{n-2;\alpha/2}S\sqrt{\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}}\right)} \tag{15.4.9a}$$

Example 15.4.2 (Steel hardness versus carbon content, revisited) *Consider the data on hardness of steel in Example 15.2.2. Find a 95% confidence interval for β_0.*

Solution: From Examples 15.2.2 and 15.3.1, we have

$$b_0 = 56.716, \quad S_{XX} = 0.01425, \quad \bar{X} = 0.2550 \quad \text{and} \quad S = 1.36$$

Since $n = 10$, $\alpha = 0.05$, and from Table A.5, we have $t_{n-2;\alpha/2} = t_{8;0.025} = 2.306$. We then have that the 95% confidence interval for β_0, from (15.4.9a), is given by

$$\left(56.716 \pm 2.306(1.36)\sqrt{\left(\frac{1}{10}+\frac{(0.2550)^2}{0.01425}\right)}\right) = (49.944, 63.488)$$

This confidence interval for β_0 is virtually identical to that obtained by using MINITAB and R (see Example 15.4.3).

Here we caution the reader that the confidence interval should not be interpreted that at $X = 0$ (i.e., without any carbon) the hardness of steel will vary between 49.944 and 63.488 psi. This is because the value $X = 0$ does not fall within the experimental region for values of X used to find the estimated regression line $\hat{Y} = b_0 + b_1X$. (More discussion on this issue is given just before Section 15.4.4.)

15.4.3 Confidence Interval for $E(Y|X)$ with Confidence Coefficient $(1 - \alpha)$

From (15.2.16) we have that for any value of X, the corresponding fitted line of $\hat{Y}$ is

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}) = b_0 + b_1 X \tag{15.4.9b}$$

Using our earlier discussion and (15.2.20), it can easily be verified that under the normality assumption $\hat{Y}$ is a random variable having the normal distribution

$$\hat{Y} \sim N\left(E(Y|X), \left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}\right]\sigma^2\right) \tag{15.4.10}$$

with $E(\hat{Y}) = E(Y|X) = \beta_0 + \beta_1 X$.

By standardizing the random variable $\hat{Y}$, we obtain

$$\frac{\hat{Y} - E(Y|X)}{\sigma\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}} \sim N(0,1) \tag{15.4.11}$$

Now, replacing σ in (15.4.11) by its estimator S, we find that

$$\frac{\hat{Y} - E(Y|X)}{S\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}} \sim t_{n-2} \tag{15.4.12}$$

From Equation (15.4.12) we have

$$P\left(\hat{Y} - t_{n-2;\alpha/2} S\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)} \leq E(Y|X) \leq \hat{Y} + t_{n-2;\alpha/2} S\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}\right) = 1 - \alpha \tag{15.4.13}$$

Thus $100(1 - \alpha)\%$ confidence interval for $E(Y|X) = \beta_0 + \beta_1 X$ is given by

$$\hat{Y} \pm t_{n-2;\alpha/2} S\sqrt{\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)} \tag{15.4.14}$$

Example 15.4.3 (Steel hardness versus carbon content data) *Consider the data on hardness of steel in Example 15.2.2. Find a 95% confidence interval for $E(Y|X = 0.25)$ by manual calculations and by using a statistical package.*

Solution: From Examples 15.4.1 and 15.4.2, we have

$$b_0 = 56.716, \quad b_1 = 90.526, \quad S_{XX} = 0.01425, \quad \bar{X} = 0.2550, \quad S = 1.36$$

and the regression line at X is such that

$$\hat{Y} = 56.716 + 90.526X$$

We note that this regression line is valid for values of X that lie in the experimental region, i.e., $0.21 \leq X \leq 0.33$. Indeed, outside this interval, $E(Y|X)$ could be quadratic in X, or exponential, etc. We further note that $X = 0.25$ lies in $[0.21,\ 0.33]$. Hence, $\hat{Y}$ at $X = 0.25$ is equal to

$$\hat{Y} = 56.716 + 90.526(0.25) = 79.3475$$

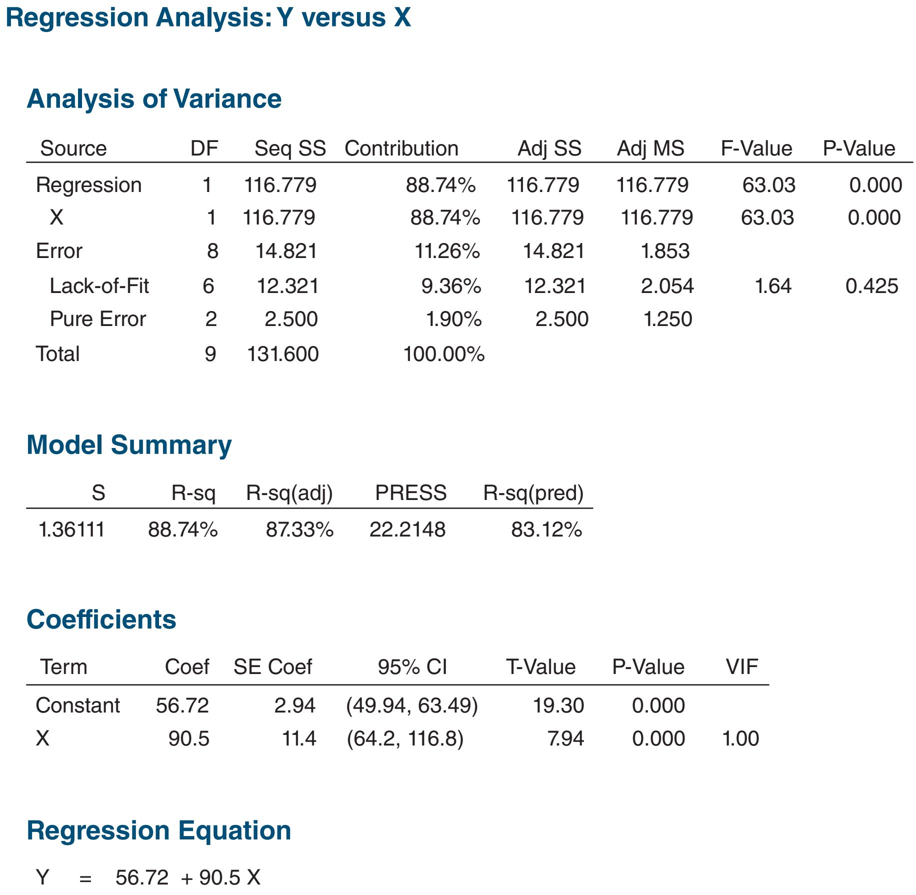
that is, the point estimate of $E(Y|X = 0.25)$ is 79.3475. Now $n = 10$, $\alpha = 0.05$, and from Table A.5 we have $t_{n-2;\alpha/2} = t_{8;0.025} = 2.306$. Hence the 95% confidence interval for $E(Y|X = 0.25)$, from (15.4.14), is given by

$$\left(79.3475 \pm (2.306)(1.36)\sqrt{\left(\frac{1}{10} + \frac{(0.25 - 0.2550)^2}{0.01425}\right)}\right) = (78.3471, 80.3479)$$

We now do Examples 15.4.1–15.4.3 by using MINITAB and R.

MINITAB

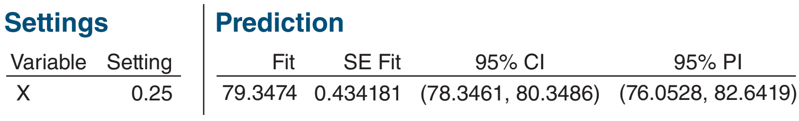
1. In a MINITAB worksheet, enter the data from Example 15.2.2 in columns C1 and C2.
2. From the Menu bar select **Stat** > **Regression** > **Regression** > **Fit Regression Model**.
3. In the dialog box that appears type C1 and C2 in the boxes below **Responses** and**Continuous predictors** (or **Categorical predictors**), respectively.
4. Select **Options** and in the resulting dialog box enter the desired confidence level (e.g. 95.0) in the box next to **Confidence level for all intervals** and select whether you want one-sided or two-sided confidence interval and click **OK**.
5. Select **Results** and check any storage options (e.g., **Method, Analysis of Variance, etc.**) and select '**Expanded tables**' option to display your results. Click **OK**, then again click **OK**. The MINITAB output will appear as shown below.
6. It includes a 95% CI for β_0 and β_1 virtually identical, except for rounding errors, to those obtained via manual calculations. Note that the confidence limits and the prediction limits appear in the **Session window**. The other portions of MINITAB output are discussed in later sections.



```
Regression Analysis: Y versus X

Analysis of Variance

Source        DF   Seq SS  Contribution   Adj SS   Adj MS  F-Value  P-Value
Regression     1  116.779        88.74%  116.779  116.779    63.03    0.000
  X            1  116.779        88.74%  116.779  116.779    63.03    0.000
Error          8   14.821        11.26%   14.821    1.853
  Lack-of-Fit  6   12.321         9.36%   12.321    2.054     1.64    0.425
  Pure Error   2    2.500         1.90%    2.500    1.250
Total          9  131.600       100.00%

Model Summary

      S     R-sq  R-sq(adj)    PRESS  R-sq(pred)
1.36111   88.74%     87.33%  22.2148      83.12%

Coefficients

Term       Coef  SE Coef          95% CI  T-Value  P-Value   VIF
Constant  56.72     2.94  (49.94, 63.49)    19.30    0.000
X          90.5     11.4   (64.2, 116.8)     7.94    0.000  1.00

Regression Equation

Y  =  56.72 + 90.5 X
```

To obtain confidence and prediction interval for $E(Y|0.25)$ take the following step.

7. From the Menu bar select **Stat** > **Regression** > **Regression** > **Predict**. Then enter the individual values or enter columns of values of X you wish to make predictions. For this example, type 0.25 on the list appears in the dialog box. Select Options and in the resulting dialog box enter the desired **Confidence level** (e.g., 95.0) and select whether you want one-sided or two-sided interval from **Type of interval** box then click **OK**, then again click **OK**.



```
Settings            | Prediction
Variable  Setting   |     Fit    SE Fit            95% CI              95% PI
X            0.25   | 79.3474  0.434181  (78.3461, 80.3486)  (76.0528, 82.6419)
```

Thus, we estimate with 95% confidence that the value of $E(Y|X = 0.25)$ is between 78.346 and 80.349.

USING R

The simple linear regression model which we discussed belongs to a class of models known as linear models. The built in 'lm()' function in R can be used to fit a wide range of such linear models. To complete the Example 15.2.2, we can use the following R code.

```
Y = c(76,79,78,86,77,80,86,79,75,82)
X = c(0.21,0.23,0.24,0.31,0.23,0.28,0.33,0.24,0.22,0.26)

#Scatter plot
plot(X,Y, main = “Scatter Plot for Data in Example 15.2.2”)

#Fitting LSR model
model = lm(Y~X)

#Summary output
model

#ANOVA output
anova(model)

#confidence intervals for regression coefficients
confint(model)

#predictions and confidence intervals for a new observation
newdata = data.frame(X = 0.25)
predict(model, newdata, interval=“confidence”)
predict(model, newdata, interval=“prediction”)
```

Thus, using the manual, MINITAB, or R results, we estimate with 95% confidence that the hardness of steel for each additional increase of 1% carbon content will increase by an amount somewhere between 642.54 and 1167.98 psi, (As noted, the units of Y are 1000 psi). Note that the values of the predictor variable X varied between 0.21 and 0.33, which is called the *experimental range*. Our estimate is best when the value of X is within the experimental range.

For prediction purposes, it is strongly recommended *not to use* a value of the predictor variable outside the experimental range (extrapolation) because, among many reasons, we do not know whether our model is valid outside of this range. For example, the relationship between the carbon content and the hardness of steel region may not be linear outside the experimental range.

15.4.4 Prediction Interval for a Future Observation Y with Confidence Coefficient $(1 - \alpha)$

Suppose that we are working with the regression model $E(Y|X) = \beta_0 + \beta_1 X$, and we find the (least-squares) regression line to be

$$\hat{Y} = b_0 + b_1 X$$

based on the data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the independent Y_i's are such that $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. Having found the regression line, we may be interested in predicting the value of a future observation, Y, to be generated at X, independent of $(X_i, Y_i), i = 1, \ldots, n$, where we assume that $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$. Of course, $E(Y|X) = \beta_0 + \beta_1 X = E(\hat{Y}|X)$, so $\hat{Y} = b_0 + b_1 X$ is a *point estimate* of the future observation Y. To find a prediction interval for Y, we consider first the random variable $Y - \hat{Y}$. We have that

$$E(Y - \hat{Y}) = E(Y) - E(\hat{Y}) = 0 \tag{15.4.15}$$

and

$$\begin{aligned} Var(Y - \hat{Y}) &= Var(Y|X) + Var(\hat{Y}|X) \\ &= \sigma^2 + \left(\tfrac{\sigma^2}{n} + \tfrac{(X-\bar{X})^2\sigma^2}{S_{XX}}\right) \end{aligned}$$

or

$$Var(Y - \hat{Y}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}\right) \tag{15.4.16}$$

Since we are assuming normality for $(Y_1, Y_2, \ldots, Y_n)$, we easily find that

$$Y - \hat{Y} \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}\right)\right) \tag{15.4.17}$$

Hence, by standardizing the random variable $(Y - \hat{Y})$, we find that

$$\frac{Y - \hat{Y}}{\sigma\sqrt{\left(1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}} \sim N(0, 1) \tag{15.4.18}$$

Now, replacing σ in (15.4.18) by its estimator S, we have

$$\frac{Y - \hat{Y}}{S\sqrt{\left(1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}} \sim t_{n-2} \tag{15.4.19}$$

From Equation (15.4.19), we have

$$P\left(-t_{n-2;\alpha/2} \le \frac{Y - \hat{Y}}{S\sqrt{\left(1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}} \le t_{n-2;\alpha/2}\right) = 1 - \alpha \tag{15.4.20}$$

so that we can write

$$\boxed{\begin{gathered} P\left(\hat{Y} - t_{n-2;\alpha/2} S\sqrt{\left(1 + \tfrac{1}{n} + \tfrac{(X-\bar{X})^2}{S_{XX}}\right)} \le Y \le \hat{Y} + t_{n-2;\alpha/2} S\sqrt{\left(1 + \tfrac{1}{n} + \tfrac{(X-\bar{X})^2}{S_{XX}}\right)}\right) \\ = 1 - \alpha \end{gathered} \qquad (15.4.21)}$$

That is, the *prediction interval* for Y to be observed at X having confidence coefficient $(1-\alpha)$ is given by

$$\left(\hat{Y} - t_{n-2;\alpha/2} S\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{S_{XX}}}, \hat{Y} + t_{n-2;\alpha/2} S\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{S_{XX}}}\right) \tag{15.4.22}$$

Example 15.4.4 (Steel hardness versus carbon content, revisited) *Consider the data on hardness of steel in Example 15.2.2. Find a 95% prediction interval for Y at $X = 0.25$.*

Solution: From Example 15.4.2, we have

$$S_{XX} = 0.01425, \quad \bar{X} = 0.2550, \quad S = 1.36, \quad n = 10, \quad 1-\alpha = 0.95, \quad \alpha/2 = 0.025$$

and from Example 15.4.3, we have that for $X = 0.25$, $\hat{Y} = 79.3475$.

Since $n = 10$, $\alpha = 0.05$, from Table A.5 we have $t_{n-2;\alpha/2} = t_{8;0.025} = 2.306$. The 95% prediction interval for Y, from (15.4.22), is given by

$$\left(79.3475 \pm (2.306)(1.36)\sqrt{\left(1+\frac{1}{10}+\frac{(0.25-0.2550)^2}{0.01425}\right)}\right) = (76.0556, 82.6394)$$

which is the 95% prediction interval for Y, to be generated at $X = 0.25$.

This prediction interval for Y is virtually identical to that obtained in Example 15.4.3 using MINITAB. Note that as expected, the prediction interval for Y with the same confidence coefficient is much wider than the confidence interval for $E(Y|X = 0.25)$.

Note: The MINITAB output also gives the value of R^2 called the *coefficient of determination*, which is a measure of how much the predictor variable X explains the linearity of the regression model. On the one hand, if all the observed values fall on the fitted line, then $R^2 = 1$; that is, the predictor variable X appearing in the model $E(Y|X) = \beta_0 + \beta_1 X$ fully explains the response variable Y. On the other hand, if $b_1 = 0$, that is, the fitted line is a horizontal line $\hat{Y} = b_0$, then the predictor variable X does not provide any information about the response variable Y, so that in this case $R^2 = 0$. In practice, it is known that

$$0 \le R^2 \le 1 \tag{15.4.23}$$

The estimate of the coefficient of determination is given by $R^2 = 1 - SSE/SST$, where $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. The value of R^2 should be used cautiously and in conjunction with the p-value for the test $H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 \neq 0$. Some general rules for using R^2 along with the p-value are

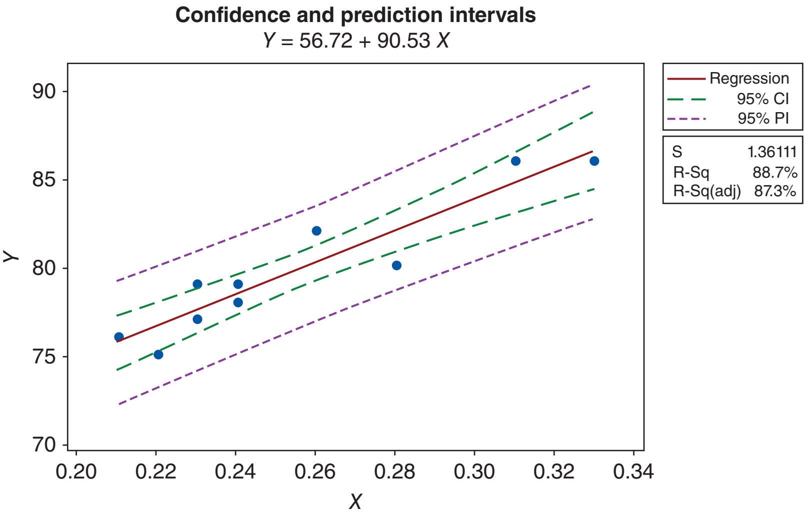


Figure 15.4.1 MINITAB printout of 95% confidence and prediction bands of regression line for the data in Example 15.2.2.

1. A high value of R^2, coupled with a very small p-value for $H_o : \beta_1 = 0$, (say) less than 5%, may indicate a strong linear relationship between the response variable and the predictor variable, provided that the residual plots indicate no abnormality.
2. A high value of R^2 and a large p-value for the test $H_0 : \beta_1 = 0$ versus $H_0 : \beta_1 \neq 0$ may indicate a strong curvilinear relationship between the response variable and the predictor variable (see Example 15.7.1).
3. A low value of R^2 and a large p-value do not necessarily preclude the existence of a relationship between the response variable and the predictor variable. Rather, this may simply mean there is a strong curvilinear relationship between them (see Example 15.7.1).

The graph in Figure 15.4.1 shows the confidence and prediction band of regression line for the data in Example 15.2.2. The graph for the confidence band is derived from the fact that a confidence interval for $E(Y|X) = \beta_0 + \beta_1 X$ is as stated in (15.4.14). If we are interested in several values of $E(Y|X)$ for different X's, then we determine the following lower and upper $100(1 - \alpha)\%$ confidence limits (see Equation (15.4.14)) for each value of X:

$$\left(b_0 + b_1 X - t_{n-2;\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}\right) \quad \text{and}$$

$$\left(b_0 + b_1 X + t_{n-2;\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}\right)$$

We plot them against X, for several values of X, and arrive at the $100(1-\alpha)\%$ confidence band for $E(Y|X)$ (see Figure 15.4.1). Similarly the prediction band in Figure 15.4.1 is based on (15.4.22). The graph of these bands may be constructed using **MINITAB** as follows:

Stat > Regression > Fitted line plot ... > Options > Display confidence and prediction intervals. Then choose an appropriate confidence level and click **OK**.

USING R

The following R code preceding the above model specification in Example 15.4.3 can be used to obtain both confidence and prediction intervals.

```
Y = c(76,79,78,86,77,80,86,79,75,82)
X = c(0.21,0.23,0.24,0.31,0.23,0.28,0.33,0.24,0.22,0.26)

#Fit LSR model
model = lm(Y~X)


#Construct CI and PI
ci.L = predict(model,data.frame(X=sort(X)), level=.95,interval=“confidence”)[,2]
ci.U = predict(model,data.frame(X=sort(X)), level=.95,interval=“confidence”)[,3]
pi.L = predict(model,data.frame(X=sort(X)), level=.95,interval=“prediction”)[,2]
pi.U = predict(model,data.frame(X=sort(X)), level=.95,interval=“prediction”)[,3]


plot(X,Y,main=“Confidence and Prediction Intervals”, col=4, pch=20, cex =2)
abline(model, lwd=2)

lines(sort(X),ci.L,lty=2, lwd=2, col=“blue”); lines(sort(X),ci.U,lty=2,lwd=2, col=“blue”)
lines(sort(X),pi.L,lty=3, lwd=2, col=“red”); lines(sort(X),pi.U,lty=3, lwd=2, col=“red”)
legend(“topleft”,c(“Fit”,“CI”,“PI”), col=c(1,4,2), lty=c(1,2,3), lwd=c(2,2,2))
```

Figure 15.4.2 shows the 95% confidence and prediction bands of the fitted regression line $\hat{Y} = 56.72 + 90.53X$ using R for the data in Example 15.2.2. This figure conveys the same information provided in MINITAB Figure 15.4.1.

PRACTICE PROBLEMS FOR SECTIONS 15.3 AND 15.4

In the following problems reference is made to the problems in Section 15.2.

1. Refer to Problem 12.
   (a) What is the estimate of σ^2?
   (b) Find 95% confidence intervals for β_0 and β_1.

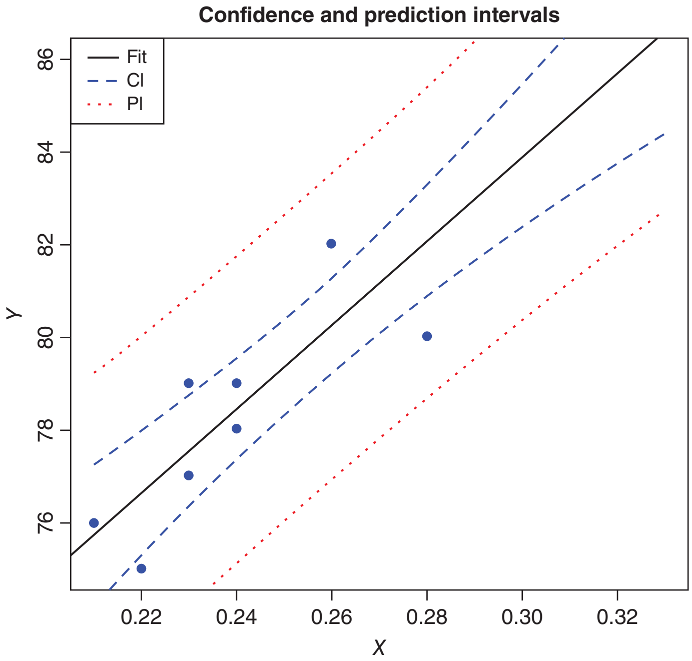


Figure 15.4.2 R output of 95% confidence and prediction bands of regression line for the data in Example 15.2.2.

(c) Find a 95% confidence interval for $E(Y|X = 3.5)$.
(d) Find a 95% prediction interval for Y when $X = 3.5$.
(e) Compare the prediction interval obtained in part (d) to the confidence interval for $E(Y|X = 3.5)$ obtained in part (c) and comment.

2. Refer to Problem 10.
   (a) What is the estimate of σ^2?
   (b) Find 95% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of % purity when $X = 1.50$, that is, $E(Y|X = 1.50)$.
   (d) Find a 95% prediction interval for purity Y when $X = 1.50$.
3. Refer to Problem 11.
   (a) What is the estimate of σ^2?
   (b) Find 95% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for $E(Y|X = 144.5)$.
   (d) Find a 95% prediction interval for Y when $X = 144.5$.
   (e) Compare the prediction interval for Y obtained in part (d) to the confidence interval for $E(Y|X = 144.5)$ obtained in part (c) and comment.

4. Refer to Problem 1.
   (a) What is the estimate of σ^2?
   (b) Find 95% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of plasma glucose when $X = 200$.
   (d) Find a 95% prediction interval for plasma glucose Y when $X = 200$.
5. Refer to Problem 2.
   (a) What is the estimate of σ^2?
   (b) Find 95% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of the chemical yield when $X = 35$.
   (d) Find a 95% prediction interval for the chemical yield Y when $X = 35$.
6. Refer to Problem 3.
   (a) What is the estimate of σ^2?
   (b) Find 99% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of shear strength when $X = 11.5$.
   (d) Find a 95% prediction interval for the shear strength Y when $X = 11.5$.
7. Refer to Problem 4.
   (a) What is the estimate of σ^2?
   (b) Find 99% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of hypertension when $X = 32$.
   (d) Find a 95% prediction interval for the hypertension Y when $X = 32$.
8. Refer to Problem 5.
   (a) What is the estimate of σ^2?
   (b) Find 99% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of rainfall when $X = 86$.
   (d) Find a 95% prediction interval for the rainfall Y when $X = 86$.
9. Refer to Problem 6.
   (a) What is the estimate of σ^2?
   (b) Find 99% confidence intervals for β_0 and β_1.
   (c) Find a 95% confidence interval for the true mean of the number of streaks when $X = 9.5$.
   (d) Find a 95% prediction interval for the number of streaks Y when $X = 9.5$.
10. Refer to Problem 7.
    (a) What is the estimate of σ^2?
    (b) Find 99% confidence intervals for β_0 and β_1.
    (c) Find a 95% confidence interval for the true mean of the OR score when $X = 85$.
    (d) Find a 95% prediction interval for the OR score Y when $X = 85$.

11. Refer to Problem 8.
    (a) What is the estimate of σ^2?
    (b) Find 95% confidence intervals for β_0 and β_1.
    (c) Find a 95% confidence interval for the true mean of the yield when $X = 16$.
    (d) Find a 95% prediction interval for the yield Y when $X = 16$.
12. Refer to Problem 9.
    (a) What is the estimate of σ^2?
    (b) Find 95% confidence intervals for β_0 and β_1.
    (c) Find a 95% confidence interval for the true mean of the tensile strength when $X = 6.5$.
    (d) Find a 95% prediction interval for the tensile strength Y when $X = 6.5$.

15.5 TESTS OF HYPOTHESES FOR β_0 AND β_1

So far in this chapter, we have discussed estimation problems for the regression coefficients, β_0 and β_1. We now focus on developing hypothesis tests concerning these regression parameters, β_0 and β_1. We again assume normality of the independent Y_i's (see Equation (15.2.4)).

15.5.1 Test of Hypotheses for β_1

To test hypotheses concerning regression parameters, we use the following procedure discussed in Chapter 9.

1. Null and alternative hypotheses: $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$
2. P(Type I error) = α
3. Test statistic: $t_{b_1} = (b_1 - \beta_{10})/(S/\sqrt{S_{XX}})$
4. Distribution of the test statistic under $H_0 :\ (b_1 - \beta_{10})/(S/\sqrt{S_{XX}}) \sim t_{n-2}$
5. Rejection or critical region: $|t_{b_1}| \geq t_{n-2;\alpha/2}$
6. If the observed value of the test statistic t_{b_1} falls in the critical region, reject the null hypothesis H_0 in favor of the alternative hypothesis H_1. Otherwise, do not reject H_0. Alternatively find the p-value. If the p-value is less than or equal to α, then reject the null hypothesis H_0 in favor of the alternative hypothesis H_1; if, however, the p-value is greater than α, then do not reject H_0.

15.5.2 Test of Hypotheses for β_0

1. Null and alternative hypotheses: $H_0 : \beta_0 = \beta_{00}$ against $H_1 : \beta_0 \neq \beta_{00}$
2. Type I error: α
3. Test statistic
$$t_{b_0} = \frac{b_0 - \beta_{00}}{S\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}}$$
4. Distribution of the test statistic under H_0
$$\frac{b_0 - \beta_{00}}{S\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim t_{n-2}$$

5. Rejection or critical region: $|t_{b_0}| \geq t_{n-2;\alpha/2}$
6. If the observed value of the test statistic t_{b_0} falls in the critical region, then reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_1. Otherwise, do not reject H_0. Alternatively, find the p-value. If the p-value is less than or equal to α, then reject the null hypothesis, H_0, in favor of the alternative hypothesis, H_1. If, however, the p-value is greater than α, then do not reject H_0.

Example 15.5.1 (Steel hardness versus carbon content) *Refer to the data on hardness of steel in Example 15.2.2. We want to test the hypothesis that the* y*-intercept is (say) 50 and also wish to test whether the slope is significantly different from 0. Thus, we test the following hypotheses at the 5% level of significance: (a)* $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$, *(b)* $H_0 : \beta_0 = 50$ *versus* $H_1 : \beta_0 \neq 50$.

Solution: (a) $n = 10$, $n - 2 = 8$; $\alpha = 0 : 05; \alpha/2 = 0.025$.

1. Hypothesis: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
2. P(Type I error) $= \alpha = 0.05$
3. Test statistic:
$$t_{b_1} = \frac{b_1 - 0}{S/\sqrt{S_{XX}}}$$
4. Distribution of the test statistic under H_0:
$$\frac{b_1 - 0}{S/\sqrt{S_{XX}}} \sim t_8$$
5. Critical region: $|t_{b_1}| > t_{8;0.025} = 2.306$
6. From Examples 15.2.2 and 15.3.1, we have $b\ _1 = 90.526$, $S = 1.36$, and $S_{XX} = 0.01425$.

Hence, the value of the test statistic under H_0 is

$$t_{b_1} = \frac{90.526}{1.36/\sqrt{0.01425}} = 7.946$$

which falls in the critical region. Thus, we reject the null hypothesis and conclude at the 5% level of significance that β_1 is significantly different from 0. Also, from the MINITAB/R outputs given earlier in Example 15.4.3, we see that the p-value is approximately 0, which leads us to the same conclusion. From the same MINITAB/R outputs we have $R^2 = 88.74\%$. After combining the p-value 0 with the high value of R^2, it seems reasonable to conclude that there is a strong linear relationship between the response variable and the predictor variable.

For part (b):

1. Hypothesis: $H_0 : \beta_0 = 50$ against $H_1 : \beta_0 \neq 50$
2. P(Type I error) $= \alpha = 0.05$
3. Test statistic:
$$t_{b_0} = \frac{b_0 - 50}{S\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}}$$

4. Distribution of the test statistic:

$$\left(\frac{b_0 - 50}{S\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \right) \sim t_8$$

5. Critical region: $|t_{b_0}| \geq t_{8;0.025} = 2.306$.
6. From Examples 15.2.2 and 15.3.1, we have $n = 10$, $b_1 = 90.526$, $S = 1.36$, and

$$E(Y|X) = \beta_0 + \beta_1 X \quad \text{with} \quad b_0 = \bar{Y} - b_1\bar{X} = 56.716.$$

Hence, the value of the test statistic is

$$t_{b_0} = \frac{56.716 - 50}{1.36\sqrt{\left(\frac{1}{10} + \frac{(0.2550)^2}{0.01425}\right)}} = \frac{6.716}{2.9368} = 2.2868$$

which does not fall in the critical region; therefore, we do not reject the null hypothesis, and we can conclude that at the 5% level of significance, β_0 is not significantly different from 50. Note that in this case, the value of the test statistic we found using MINITAB and R is entirely different from what we found above. This is because the test statistic in the output of MINITAB is based on the null hypothesis ($H_0 : \beta_0 = 0$).

In order to test a different hypothesis, using MINITAB or R, we do not have a built-in procedure, but we can proceed as follows. Suppose that we need to find a confidence interval for the predicted value at $X = 0$ and to check if the value $\beta_0 = 50$, under the null hypothesis, falls in the confidence interval. If it does fall in the confidence interval, then we do not reject the null hypothesis. Otherwise, we reject it. In the present example, from the R output, the 95% confidence interval is (49.938, 63.494), which contains the value specified by the null hypothesis, $\beta_0 = 50$. Also, the confidence interval obtained by using MINITAB contains the value $\beta_0 = 50$. Therefore, we do not reject the null hypothesis at significance level 0.05.

Example 15.5.2 (Percentage of waste solids removed from a filtration system) *A study was instituted to determine the percent of waste solids removed from a filtration system as a function of the flow rate of the effluent being fed into the system. It was decided to use flow rates* X *of 2, 4, ..., 14 gal/min and to observe* Y, *the percent of waste solid removed, when each of these flow rates was used. The study yielded the data in Table 15.5.1.*

Using MINITAB or R do the following:

(a) Construct a scatter plot for the data in Table 15.5.1.
(b) Fit the regression line and test hypotheses $\beta_1 = 0$ versus $\beta_1 \neq 0$, *and* $\beta_0 = 0$ versus $\beta_0 \neq 0$.

Table 15.5.1 Flow rate and percent of waste solid removed.

Y	24.3	19.7	17.8	14.0	12.3	7.2	5.5
X	2	4	6	8	10	12	14

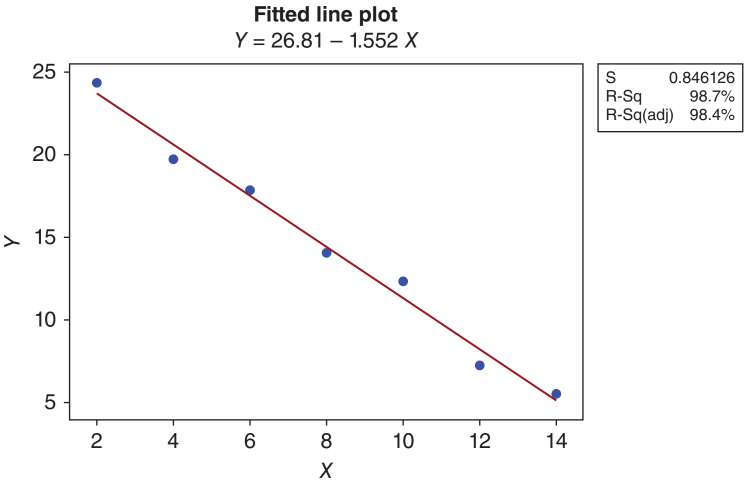


Figure 15.5.1 Scatter plot for the data in Table 15.5.1.

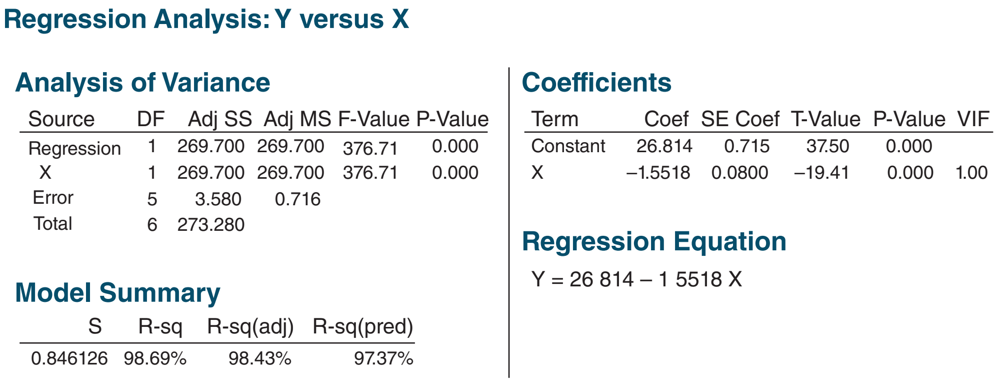
(c) *Construct a graph showing confidence and prediction bands for the fitted regression line.*

Solution: MINITAB

Using the same steps as given in Examples 2.9.1 and 15.4.3, we have

(a) Plot of the data in Table 15.5.1 and the fitted regression line as shown in Figure 15.5.1.
(b) The regression equation takes the form $\hat{Y} = b_0 + b_1 X = 26.814 - 1.5518X$ (see Figure 15.5.1)

The summary of the regression analysis (From the Menu bar select **Stat** > **Regression** > **Regression** > **Fit Regression Model...**).



```
Regression Analysis: Y versus X

Analysis of Variance                                  Coefficients
Source      DF  Adj SS   Adj MS  F-Value P-Value      Term       Coef    SE Coef  T-Value  P-Value  VIF
Regression   1  269.700  269.700  376.71   0.000      Constant   26.814   0.715    37.50    0.000
  X          1  269.700  269.700  376.71   0.000      X         –1.5518   0.0800  –19.41    0.000   1.00
 Error       5    3.580    0.716
 Total       6  273.280                               Regression Equation
                                                      Y = 26 814 – 1 5518 X
Model Summary
        S     R-sq   R-sq(adj)   R-sq(pred)
 0.846126   98.69%     98.43%       97.37%
```

Since the p-values for testing both β_0 and β_1 are approximately 0, we reject the null hypotheses that $\beta_0 = 0$ and that $\beta_1 = 0$. Moreover, because the value of R^2 is very high, in conjunction with the p-value ≈ 0, we can conclude that there is a strong linear relationship between percentage of waste solid removed and the flow rate.

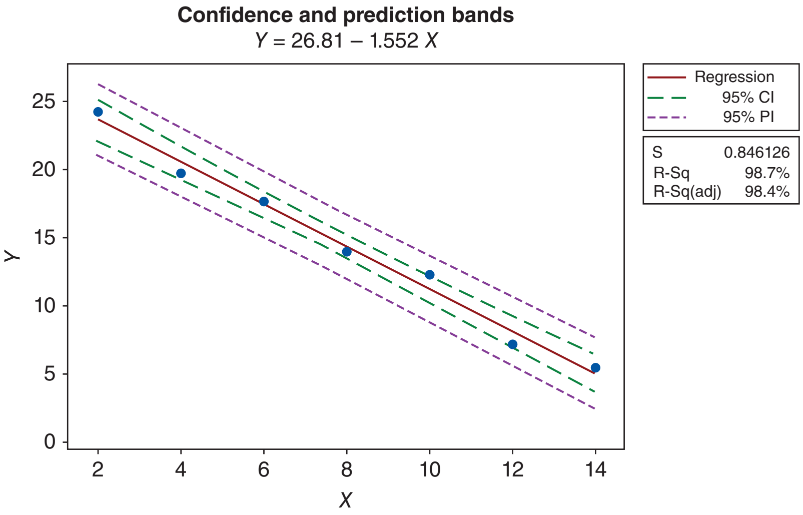


Figure 15.5.2 MINITAB printout of the confidence and prediction bands of regression line for the data in Table 15.5.1.

(c) Figure 15.5.2 shows the 95% confidence and prediction bands of the fitted regression line ($\hat{Y} = b_0 + b_1X = 26.814 - 1.5518X$) in Part (b).

USING R

Following the steps outlined for Examples 15.4.3 and 15.4.4, we get results virtually identical to the ones obtained using MINITAB. Use the following R code.

```
Y = c(24.3,19.7,17.8,14.0,12.3,7.2,5.5)
X = c(2,4,6,8,10,12,14)

#Fit LSR model
model = lm(Y ~ X)
summary(model)


#Scatter plot
plot(X,Y, col=4, pch=20, cex =2,main = “Scatter Plot for Data in Table 15.5.1”)
abline(model, lwd=2)


#Calculate and plot CI and PI
ci.L = predict(model,data.frame(X=sort(X)), level=.95,interval=“confidence”)[,2]
ci.U = predict(model,data.frame(X=sort(X)), level=.95,interval=“confidence”)[,3]
pi.L = predict(model,data.frame(X=sort(X)), level=.95,interval=“prediction”)[,2]
pi.U = predict(model,data.frame(X=sort(X)), level=.95,interval=“prediction”)[,3]


plot(X,Y,main=“Confidence and Prediction Intervals”, col=4, pch=20, cex =2)

abline(model, lwd=2)
lines(sort(X),ci.L,lty=2, lwd=2, col=“blue”); lines(sort(X),ci.U,lty=2,lwd=2, col=“blue”)
lines(sort(X),pi.L,lty=3, lwd=2, col=“red”); lines(sort(X),pi.U,lty=3, lwd=2, col=“red”)
legend(“topright”,c(“Fit”,“CI”,“PI”), col=c(1,4,2), lty=c(1,2,3), lwd=c(2,2,2))
```

We now give a summary in Table 15.5.2 of the formulas derived so far in this chapter. These formulas come in handy for the many computations in a regression analysis for the model $E(Y|X) = \beta_0 + \beta_1 X$.

Example 15.5.3 (Percentage of waste solids removed from a filtration system) *Using the formulas for confidence intervals presented in Table 15.5.2 for the data in Example 15.5.2, find 95% confidence intervals for* $\beta_0, \beta_1, E(Y|X)$, *and* Y *at* X = *4.*

Solution: Using the data in Example 15.5.2 and the formulas given in Table 15.5.2, we have

$$\bar{X} = 8.0, \quad \bar{Y} = 14.4, \quad S_{XX} = 112, \quad S_{YY} = 273.28, \quad S_{XY} = -173.8, \quad n = 7,$$
$$b_1 = -1.55, \quad b_0 = 26.81$$

Substituting the values in appropriate formulas in Table 15.5.2, we have, for $1 - \alpha = 0.95$, that, as the reader may verify, 95% confidence intervals are

$$\text{For } \beta_1: \left(-1.55 \pm 2.571\sqrt{\frac{1}{112}(0.72)}\right) = (-1.55 \pm 0.21) = (-1.76, -1.34)$$

$$\text{For } \beta_0: \left(26.81 \pm 2.571\sqrt{\left[\frac{1}{7} + \frac{(8.0)^2}{112}\right](0.72)}\right) = (26.81 \pm 1.84) = (24.97, 28.65)$$

Table 15.5.2 Certain formulas useful for computations in regression analysis.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

$$S_{XX} = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}, \quad S_{YY} = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

$$S_{XY} = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}$$

$$b_1 = S_{XY}/S_{XX}, \quad b_0 = \bar{Y} - b_1\bar{X}$$

$$SSE = S_{YY} - b_1, \quad S_{XY} = S_{YY} - [(S_{XY})^2/S_{XX}]$$

$$\hat{\sigma}^2 = S^2 = MSE = SSE/(n-2)$$

Confidence intervals with confidence coefficient $(1 - \alpha)$

$$\text{For } \beta_1: \left(b_1 \pm t_{n-2;\alpha/2}\frac{S}{\sqrt{S_{XX}}}\right)$$

$$\text{For } \beta_0: \left(b_0 \pm t_{n-2;\alpha/2} S\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}\right)$$

$$\text{For } E(Y|X): \left(\hat{Y} \pm t_{n-2;\alpha/2} S\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}}\right)$$

$$\text{For } (Y|X): \left(\hat{Y} \pm t_{n-2;\alpha/2} S\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}}\right), \ \hat{Y} = b_0 + b_1 X$$

$$\text{For } E(Y|X=4):\left(20.61 \pm 2.571\sqrt{\left[\frac{1}{7}+\frac{(4-8.0)^2}{112}\right](0.72)}\right) = (20.61 \pm 1.16)$$
$$= (19.45, 21.77)$$
$$\text{For } (Y|X=4):\left(20.61 \pm 2.571\sqrt{\left[1+\frac{1}{7}+\frac{(4-8.0)^2}{112}\right](0.72)}\right) = (20.61 \pm 2.47)$$
$$= (18.14, 23.08)$$

PRACTICE PROBLEMS FOR SECTION 15.5

In the following problems, reference is made to the problems in Section 15.2. Whenever possible, to perform the testing of hypotheses in the following problems, use the confidence intervals obtained in the practice problems for Sections 15.3 and 15.4 by selecting the appropriate size of the level of significance. In each problem state the level of significance you use. Further, in each problem, state your conclusions when you reject or do not reject either of the hypotheses.

1. Refer to Problem 1. Test the following hypotheses using $\alpha = 0.05$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \text{ versus } \quad H_1: \beta_0 \neq 0$$
2. Refer to Problem 2. Test the following hypotheses using $\alpha = 0.05$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$
3. Refer to Problem 3. Test the following hypotheses using $\alpha = 0.01$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$
4. Refer to Problem 4. Test the following hypotheses using $\alpha = 0.01$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$
5. Refer to Problem 5. Test the following hypotheses using $\alpha = 0.01$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$
6. Refer to Problem 6. Test the following hypotheses using $\alpha = 0.01$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$
7. Refer to Problem 7. Test the following hypotheses using $\alpha = 0.01$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$
8. Refer to Problem 8. Test the following hypotheses using $\alpha = 0.05$:
$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0 \quad \text{and} \quad H_0: \beta_0 = 0 \quad \text{versus} \quad H_1: \beta_0 \neq 0$$

9. Refer to Problem 9. Test the following hypotheses using $\alpha = 0.05$:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0 \quad \text{and} \quad H_0 : \beta_0 = 0 \quad \text{versus} \quad H_1 : \beta_0 \neq 0$$

10. Refer to Problem 10. Test the following hypotheses using $\alpha = 0.05$:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0 \quad \text{and} \quad H_0 : \beta_0 = 0 \quad \text{versus} \quad H_1 : \beta_0 \neq 0$$

11. Refer to Problem 11. Test the following hypotheses using $\alpha = 0.05$:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0 \quad \text{and} \quad H_0 : \beta_0 = 0 \quad \text{versus} \quad H_1 : \beta_0 \neq 0$$

12. Refer to Problem 12. Test the following hypotheses using $\alpha = 0.05$:

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0 \quad \text{and} \quad H_0 : \beta_0 = 0 \quad \text{versus} \quad H_1 : \beta_0 \neq 0$$

15.6 ANALYSIS OF VARIANCE APPROACH TO SIMPLE LINEAR REGRESSION ANALYSIS

The *analysis of variance* approach to simple regression analysis is just another technique for the various problems we discussed in this chapter. Before we present this technique in more detail, we need to define certain terms commonly used in an analysis of variance.

Uncorrected or crude total sum of squares: $\sum_{i=1}^{n} Y_i^2$
Corrected total sum of Squares: $SS_{Total} = S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$
Correction factor: $n\bar{Y}^2$
Sum of squares due to regression: $SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
Error or residual sum of squares: $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = S_{YY} - b_1 S_{XY}$

In the analysis of variance technique, we partition the corrected total sum squares into two sum of squares, that is *sum of squares due to regression* and *residual sum of squares* as follows.

From (15.2.16), we have that

$$\hat{Y}_i - \bar{Y} = b_1(X_i - \bar{X}) \tag{15.6.1}$$

Squaring both sides of (15.6.1) yields, as the reader can verify,

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = b_1 S_{XY} \tag{15.6.2}$$

Now from (15.3.8), we have

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = S_{YY} - b_1 S_{XY} \tag{15.6.3}$$

From (15.6.2) and (15.6.3) we then have

$$S_{YY} = SS_{Total} = SSE + SSR \tag{15.6.4}$$

In Chapter 17, we will show that under the assumption that ε_i's are independent $N(0, \sigma^2)$ random variables, SSE and SSR are independent with probability distributions given by

$$SSE \sim \sigma^2 \chi^2_{n-2} \tag{15.6.5}$$

$$SSR \sim \sigma^2 \chi^2_1(\lambda) \tag{15.6.6}$$

Here λ denotes the parameter of noncentrality and is given by

$$\lambda = \beta_1^2 S_{XX} \tag{15.6.7}$$

while $\chi^2_m(\lambda)$ denotes a random variable that is distributed as a noncentral chi-square random variable, with m degrees of freedom and with noncentrality parameter λ.

The results shown in (15.6.2) through (15.6.7) are often summarized in a table, called an *analysis of variance table*, or simply an ANOVA table, as in Table 15.6.1. *An Analysis of Variance Table or ANOVA Table may be described briefly as a table giving the breakdown of the total sum of squares and corresponding degrees of freedom into various components of sums of squares and their related degrees of freedom.*

Note: If $\beta_1 = 0$, then from (15.6.7) $\lambda = 0$, so that $E(MSR) = E(MSE) = \sigma^2$.

Table 15.6.1 Analysis of variance for a fitted straight line.

Source of Variation	Sum of squares	Degrees of Freedom	Mean square	F-Ratio	Expected mean square
Regression	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \dfrac{SSR}{1}$	$\dfrac{MSR}{MSE}$	$\sigma^2 + \lambda$
Residual	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \dfrac{SSE}{n-2}$		σ^2
Total	$SS_{Total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$n-1$			

Example 15.6.1 (Percentage of waste solids data) *Refer to the data on percentage of waste solid removed given in Table 15.5.1 and used in Example 15.5.2. Construct an analysis of variance table and test the hypothesis*

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

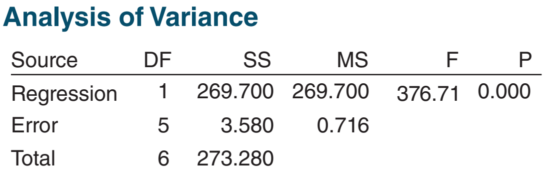
Solution: From the data in Table 15.5.1, we obtain

$$n = 7,\ \ \bar{X} = 8.0,\ \ \bar{Y} = 14.4,\ \ S_{XX} = 112,\ \ S_{YY} = 273.28,\ \ S_{XY} = -173.8,$$
$$n - 2 = 5$$
$$b_1 = S_{XY}/S_{XX} = -173.8/112 = -1.5518$$
$$SSR = b_1 S_{XY} = (-1.5518) \times (-173.8) = 269.70$$
$$SS_{Total} = S_{YY} = 273.28$$
$$SSE = SS_{Total} - SSR = 293.28 - 269.70 = 3.58$$

Thus the analysis of variance table for the data in Table 15.5.1 is as shown in Table 15.6.2.

The analysis of variance table is used to test the hypothesis $H_0 : \beta_1 = 0$ against the two-sided alternative $H_1 : \beta_1 \neq 0$ by computing F, the ratio of the mean square due to regression to the residual mean square, and comparing this value to the tabular value $F_{1,n-2}; \alpha$ of Snedecor's $F_{1,n-2}$ -distribution. If the computed ratio is greater than the upper $100\alpha\%$ point of the Snedecor's $F_{1,n-2}$ -distribution, then we reject the null hypothesis. Otherwise, we do not reject the null hypothesis. In this example, the ratio of the mean square due to regression to the residual mean square is $269.7/0.716 = 376.68$, which is greater than the value of $F_{1,5;0.05} = 6.61$. Thus, we reject the null hypothesis $H_0 : \beta_1 = 0$ at the 5% level of significance.

Note that when using MINITAB, the output automatically gives the ANOVA table. The MINITAB output for the data in Table 15.5.1 produces the following ANOVA table (DF = degrees of freedom):



```
Analysis of Variance

Source      DF       SS        MS        F      P
Regression   1   269.700   269.700   376.71  0.000
Error        5     3.580     0.716
Total        6   273.280
```

Furthermore, the value of R^2 can be found using the ANOVA table as follows:

$$R^2 = \frac{SSR}{SS_{Total}} = \frac{269.70}{273.28} = 0.9869$$

which matches the value found in Example 15.5.2.

Table 15.6.2 Analysis of variance table for the data in Table 15.5.1.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F -Ratio
Regression	269.70	1	269.70	$F = 376.68$
Residual	3.58	5	0.716	
Total	273.28	6		

Example 15.6.2 (Shell production in an ammunition plant) *Consider the following ammunition plant data comparing ambient temperature and shell production. Using MINITAB and R, construct an analysis of variance table and test the hypothesis*

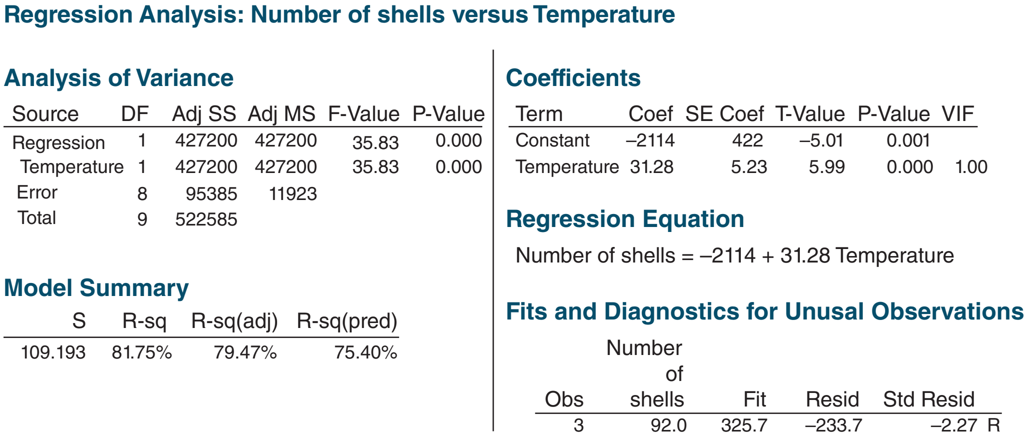
$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0$$

Temperature, X	80	68	78	79	87	74	86	92	77	84
Number of shells, Y	411	29	92	425	618	343	604	752	192	573

Solution:

MINITAB

1. In a MINITAB worksheet, enter the data above in columns C1 and C2.
2. From the main toolbar, select **Stat** > **Regression** > **Regression** > **Fit Regression Model**.
3. In the dialog box that appears type C1 (Temperature) and C2 (Number of shells) in the boxes next to **Response** and **Predictors**, respectively. Then click **OK**. Portions of the MINITAB output appear as follows:



```
Regression Analysis: Number of shells versus Temperature

Analysis of Variance                                 Coefficients
Source        DF  Adj SS  Adj MS  F-Value  P-Value   Term         Coef  SE Coef  T-Value  P-Value   VIF
Regression     1  427200  427200    35.83    0.000   Constant    –2114      422    –5.01    0.001
  Temperature  1  427200  427200    35.83    0.000   Temperature 31.28     5.23     5.99    0.000  1.00
 Error         8   95385   11923
 Total         9  522585                             Regression Equation
                                                     Number of shells = –2114 + 31.28 Temperature
Model Summary
       S    R-sq  R-sq(adj)  R-sq(pred)              Fits and Diagnostics for Unusal Observations
 109.193  81.75%     79.47%      75.40%                        Number
                                                                   of
                                                       Obs     shells    Fit   Resid  Std Resid
                                                         3       92.0  325.7  –233.7     –2.27  R
```

USING R

From the data above, we use temperature as the predictor and the number of shells as the response. Use the following R code to get the subsequent ANOVA output.

```
Y = c(411,29,92,425,618,343,604,752,192,573)
X = c(80,68,78,79,87,74,86,92,77,84)

#Fitting LSR model
model = lm(Y~ X)


#ANOVA and Summary outputs
anova(model)
summary(model)
```

From both the MINITAB and R output, we find a relatively large F-value with very small p-value (≈ 0.000). Thus, we have ample evidence at the 0.05 or any other level of significance to reject the null hypothesis that $\beta_1 = 0$.

Notes:

1. The rationale for using the ratio of the mean square due to regression to the residual mean square to test the hypothesis, $H_0 : \beta_1 = 0$ against the two-sided alternative $H_1 : \beta_1 \neq 0$, is as follows. Remember, under the null hypothesis, $\lambda = \beta_1^2 S_{XX} = 0$. Hence, from the expected mean square column in Table 15.6.1, it follows that when $\lambda = 0$, both $MSR = SSR/1$ and $MSE = SSE/(n - 2)$ are unbiased estimators of σ^2. This implies that the observed values of MSR and MSE should not differ significantly under $H\ _0$. Furthermore, from Equations (15.6.5) and (15.6.6), we have that the ratio of MSR to MSE under $H\ _0$ is distributed as a Snedecor's F-distribution, so the observed values of MSR and MSE do not differ significantly at the α level of significance if the observed ratio of MSR to MSE is less than the tabular value of $F_{1,n-2;\alpha}$. In other words, we do not reject the null hypothesis $H_0 : \beta_1 = 0$ if the observed ratio of MSR to MSE is less than $F_{1,n-2;\alpha}$.
2. Recall from Examples 15.6.1 and 15.5.2 that on the basis of two-sided t test, the data give evidence to support the hypothesis $H_1 : \beta_1 \neq 0$. In fact, the observed $t_5 = -19.41$. Recall from Section 7.3 that $t_{df}^2 \sim F_{1,df}$, so the observed $F = (-19.4091)^2 = 376.74$, which agrees with the entry for the F-ratio in Table 15.6.2, except for some rounding error.

PRACTICE PROBLEMS FOR SECTION 15.6

In the following problems, reference is made to the problems in Section 15.2.

1. Construct the ANOVA table for the data in Problem 12 and use results in this table to evaluate the fitted model in Problem 12. Use $\alpha = 0.05$.
2. Refer to Problem 10.
   (a) Construct the ANOVA table for the data in Problem 10.
   (b) Use the ANOVA table in part (a) to evaluate the fitted model in Problem 10. Use $\alpha = 0.01$.
   (c) Calculate the coefficient of determination R^2.
3. Refer to Problem 11.
   (a) Construct the ANOVA table for the data in Problem 11.
   (b) Use the ANOVA table in (a) to evaluate the fitted model in Problem 11. Use $\alpha = 0.01$.
   (c) Determine the observed level of significance (p-value) in (b).
   (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and of the p-value determined in (c).
4. The following data give the methyl mercury intake X and whole blood mercury Y in 12 subjects exposed to methyl mercury by eating contaminated fish. Assume

that the simple linear regression is a suitable model to describe the following data (from Daniel, 2006, used with permission).

X	180	200	230	410	600	550	275	580	105	250	460	650
Y	90	120	125	290	310	290	170	375	70	105	205	480

(a) Construct ANOVA table for these data.
(b) Test the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Use $\alpha = 0.01$.
(c) Determine the observed level of significance (p-value) in (b).
(d) Determine the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and the p-value determined in (c).

5. Refer to Problem 1.
   (a) Construct the ANOVA table for the data in Problem 1.
   (b) Use the ANOVA table in part (a) to evaluate the fitted model in Problem 1. Use $\alpha = 0.05$.
   (c) Determine the observed level of significance (p-value) in (b).
   (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and the p-value determined in (c).
6. Refer to Problem 2.
   (a) Construct the ANOVA table for the data in Problem 2.
   (b) Use the ANOVA table in (a) to evaluate the fitted model in Problem 2. Use $\alpha = 0.05$.
   (c) Determine the observed level of significance (p-value) in (b).
   (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and of the p-value determined in (c).
7. Refer to Problem 3.
   (a) Construct the ANOVA table for the data in Problem 3.
   (b) Use the ANOVA table in part (a) to evaluate the fitted model in Problem 3. Use $\alpha = 0.05$.
   (c) Determine the observed level of significance (p-value) in (b).
   (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and of the p-value determined in (c).
8. Refer to Problem 7.
   (a) Construct the ANOVA table for the data in Problem 7.
   (b) Use the ANOVA table in (a) to evaluate the fitted model in Problem 7. Use $\alpha = 0.01$.
   (c) Determine the observed level of significance (p-value) in (b).
   (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and of the p-value determined in (c).
9. Refer to Problem 8.
   (a) Construct the ANOVA table for the data in Problem 8.
   (b) Use the ANOVA table in part (a) to evaluate the fitted model in Problem 8. Use $\alpha = 0.01$.
   (c) Determine the observed level of significance (p-value) in (b).
   (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and of the p-value determined in (c).

10. Refer to Problem 9.
    (a) Construct the ANOVA table for the data in Problem 9.
    (b) Use the ANOVA table in (a) to evaluate the fitted model in Problem 9. Use $\alpha = 0.01$.
    (c) Determine the observed level of significance (p-value) in (b).
    (d) Calculate the coefficient of determination R^2. Give the practical interpretation of the value of R^2 and of the p-value determined in (c).

15.7 RESIDUAL ANALYSIS

So far in this chapter, we have discussed the problems of estimation and hypothesis testing for the regression parameters in the simple linear regression model (15.2.4) without paying much attention to the validity of our assumptions about the model. If any of the assumptions about the model are violated, that is, if the assumptions of independence (random errors), normality of ε_i (random errors), or constant variance are violated, then any conclusions we make about the data conforming to the model (15.2.4) may be invalid. In this section, we check the assumptions of the model by studying the observed residuals using some graphical techniques. This study of observed residuals is usually called a *residual analysis.* The residual analysis, in addition, gives us information about other departures from the model, such as the presence of outliers in the data, the omission of some important independent variables, and/or quadratic terms of the independent variables.

We present several graphs of residuals for the data in Example 15.2.2 that give some insight about the validity of the assumptions in model (15.2.4) or any other departures from the simple linear regression model. These graphs include a run chart, box plot, normal probability plot, plot of residuals versus predictor variable, or plot of residuals versus fitted values (Table 15.3.1 gives Y_i's and corresponding residuals $Y_i - \hat{Y}_i$).

The various plots in Figure 15.7.1 show that the model (15.2.4) is quite appropriate for the data on hardness of steel. An interpretation of plots (a) to (e) in Figure 15.7.1 is as follows:

1. Since all the points almost fall on a straight-line, the residuals are normally distributed.
2. Since all the points are randomly scattered and fall within a rectangular band, the variance seems fairly constant.
3. The plot of residuals versus the observation order does not present any patterns, which could violate the assumption that the ε_i's are independent.
4. The box plot shows no outliers, which implies that there are no unusual observations. Plot (e) provides the same information as plot (b).

Some other *typical* plots that may arise in residual analysis and that indicate departures from the linear regression model (15.2.4), are shown in Figure 15.7.2.

The scatter plot of residuals versus the predictor variable in Figure 15.7.2a has a curved pattern: for some smaller and larger predictor values the residuals are negative, while for intermediate values, the residuals are positive. This indicates that a linear regression model is not appropriate and that, likely, a quadratic term of the predictor variable should be included in the model.

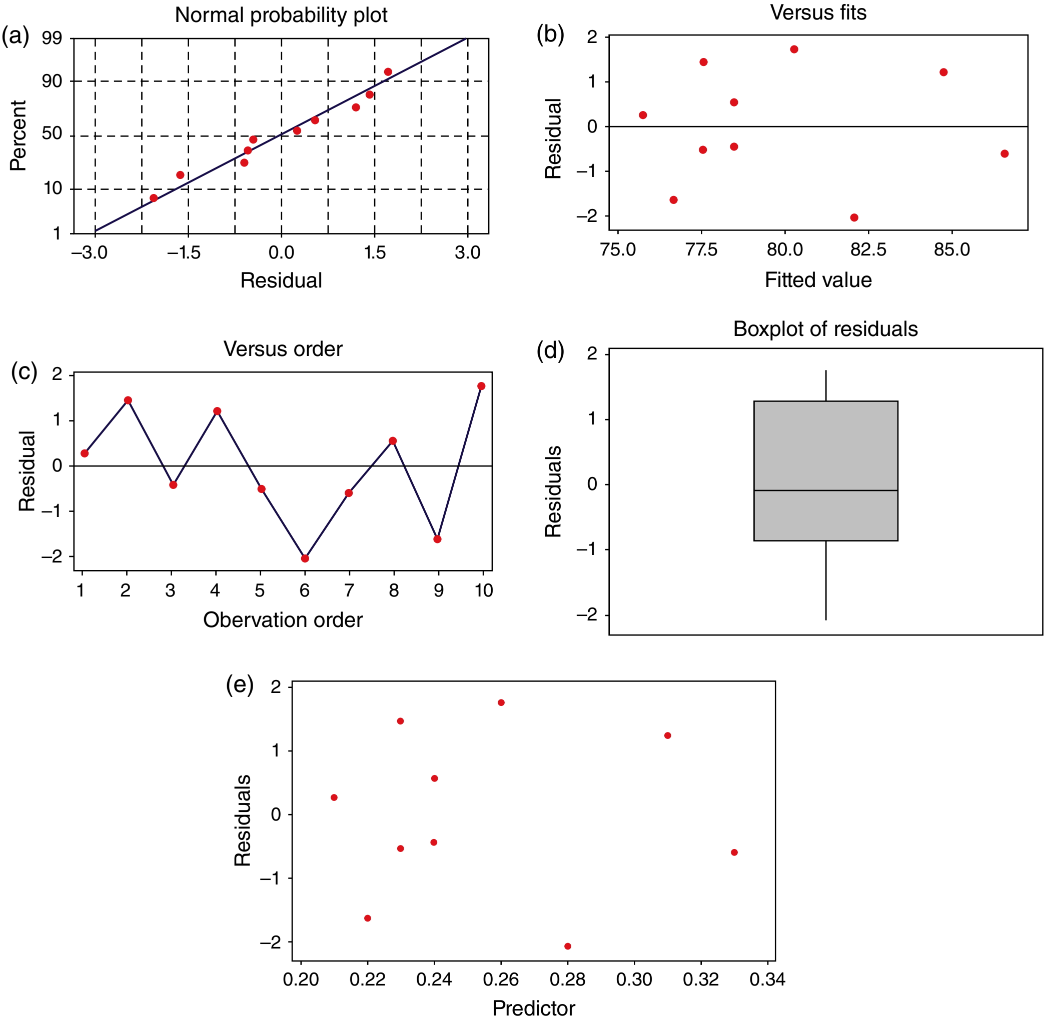


Figure 15.7.1 MINITAB plots of the residuals for the data in Example 15.5.2 on hardness of steel.

The scatter plot of residuals versus the predictor variable in Figure 15.7.2b shows that the dispersion of residuals is increasing as the value of the predictor variable increases. This indicates that the assumption of constant variance is not valid. In order to validate this assumption, we would need to use some data transformation on the response or predictor variable that may help stabilize the variance.

The normality plot of residuals in Figure 15.7.2d shows that the normality condition of the model (15.2.4) is also violated. Note that MINITAB also provides the p-value for testing the null hypothesis H_0 : Residuals are normally distributed versus H_1 : Residual are not normally distributed. In Figure 15.7.2d, we note that the p-value is 0.012 (not shown in the diagram), so we can reject the null hypothesis at any level of significance greater than the p-value 0.012. Further, we note that one of the residuals is quite a bit off the straight-line, which indicates that the observation corresponding to this residual is an unusual one.

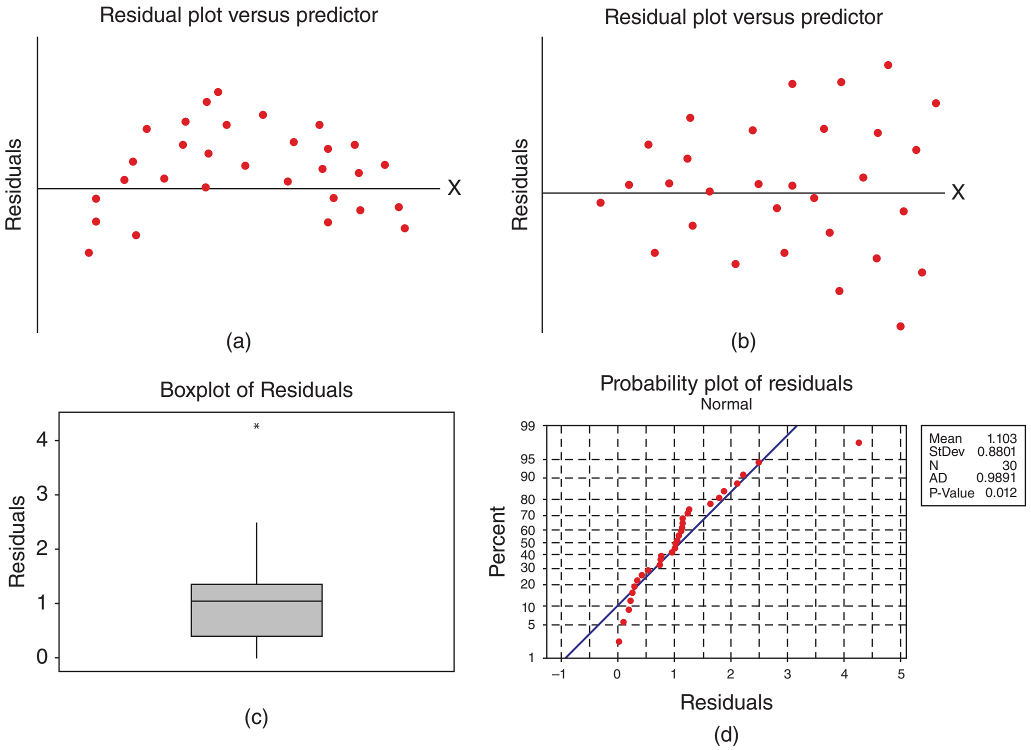


Figure 15.7.2 Some other typical plots in residual analysis, showing departure from the simple linear regression model (15.2.4).

This assertion is also confirmed by the box plot in Figure 15.7.2c (Figures 15.7.2c and 15.7.2d represent the same set of residuals). The plots in Figure 15.7.2 leads to the conclusion that various assumptions of the simple linear regression model are violated. Certain remedies are available to validate some of these assumptions. These remedies include some transformation of variables; as mentioned above, some discussion of this topic is presented in Section 15.8.

Example 15.7.1 (Amount of phosphate versus soybean yield) *An experiment is conducted to determine the amount of phosphate needed per acre to optimize the yield of a soybean crop when it is known how much potassium and lime is needed. The data in Table 15.7.1 provide the necessary information. Use one of the software packages to fit an appropriate model to these data.*

Table 15.7.1 Data on soybean crop experiment.

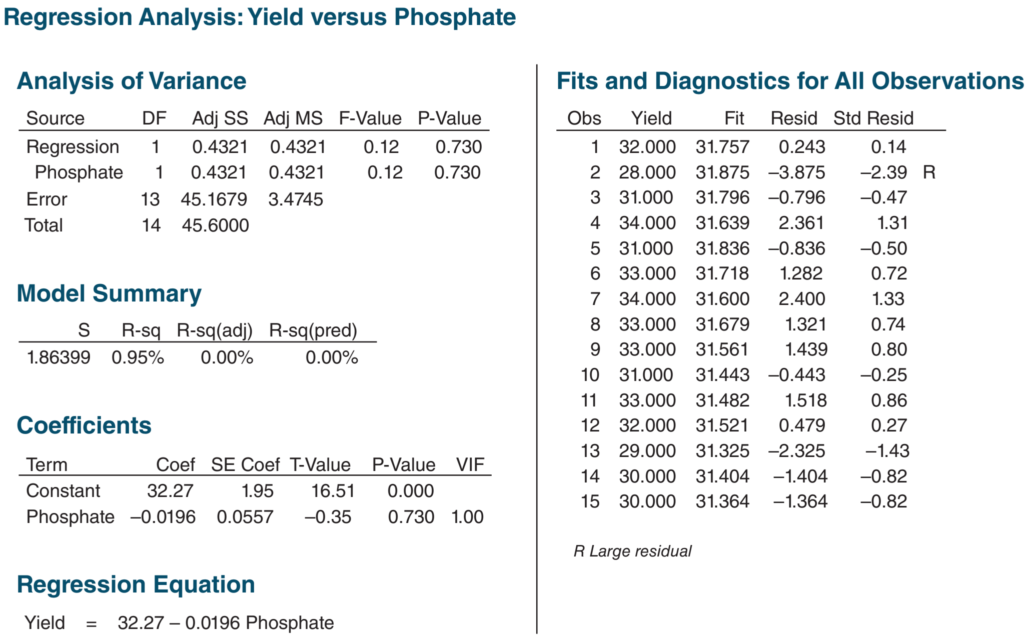
Yield Y bushels	32	28	31	34	31	33	34	33	33	31	33	32	29	30	30
Phosphate X lb	26	20	24	32	22	28	34	30	36	42	40	38	48	44	46

Solution:

MINITAB

When fitting a linear regression model, it is generally recommended to first plot the data in a scatter plot to visualize the trend presence in the data before fitting a model. For the data in this example, the following scatter plot in Figure 15.7.3 is obtained. It is clear from the scatter plot that a simple linear regression model is not adequate for the data available in this study. However, to understand the danger of fitting a straight-line through this data we will proceed with fitting the simple linear regression model.

From the Menu bar select **Stat** > **Regression** > **Regression**; then enter yields in the box next to **Response** and phosphate in the box next to **Predictors**. Then select **Results** (select Fits and diagnostics: For all observations), **Graphs**, and **options** one by one, and check the entries corresponding to the results you desire to have. Then click **OK**. The following result appears in the session window. *Note*: the number of classes in histogram can be changed by double clicking on histogram bars and selecting Number of intervals to be 5 in the Binning option in the Edit Bars menu.



```
Regression Analysis: Yield versus Phosphate

Analysis of Variance

Source      DF   Adj SS   Adj MS  F-Value  P-Value
Regression   1   0.4321   0.4321     0.12    0.730
  Phosphate  1   0.4321   0.4321     0.12    0.730
Error       13  45.1679   3.4745
Total       14  45.6000

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
1.86399   0.95%      0.00%       0.00%

Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant     32.27     1.95    16.51    0.000
Phosphate  -0.0196   0.0557    -0.35    0.730  1.00

Regression Equation

Yield  =  32.27 - 0.0196 Phosphate
```

```
Fits and Diagnostics for All Observations

Obs   Yield     Fit   Resid  Std Resid
  1  32.000  31.757   0.243       0.14
  2  28.000  31.875  -3.875      -2.39  R
  3  31.000  31.796  -0.796      -0.47
  4  34.000  31.639   2.361       1.31
  5  31.000  31.836  -0.836      -0.50
  6  33.000  31.718   1.282       0.72
  7  34.000  31.600   2.400       1.33
  8  33.000  31.679   1.321       0.74
  9  33.000  31.561   1.439       0.80
 10  31.000  31.443  -0.443      -0.25
 11  33.000  31.482   1.518       0.86
 12  32.000  31.521   0.479       0.27
 13  29.000  31.325  -2.325      -1.43
 14  30.000  31.404  -1.404      -0.82
 15  30.000  31.364  -1.364      -0.82

R Large residual
```

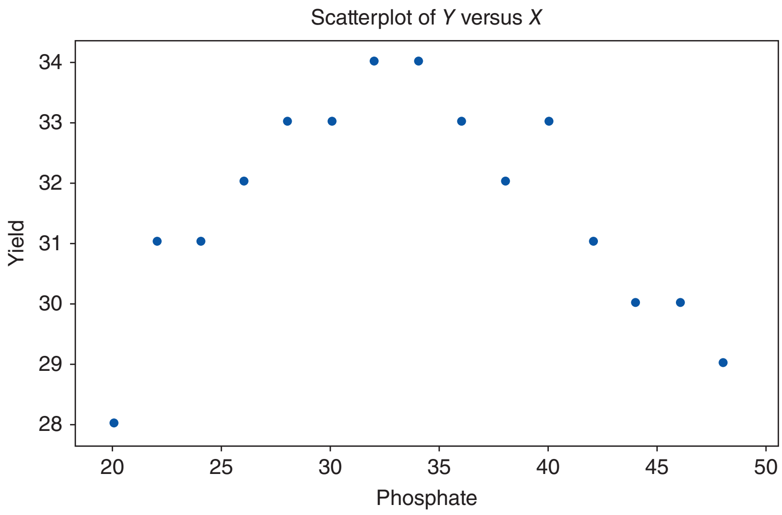


Figure 15.7.3 MINITAB scatter plot for the data in Example 15.7.1.

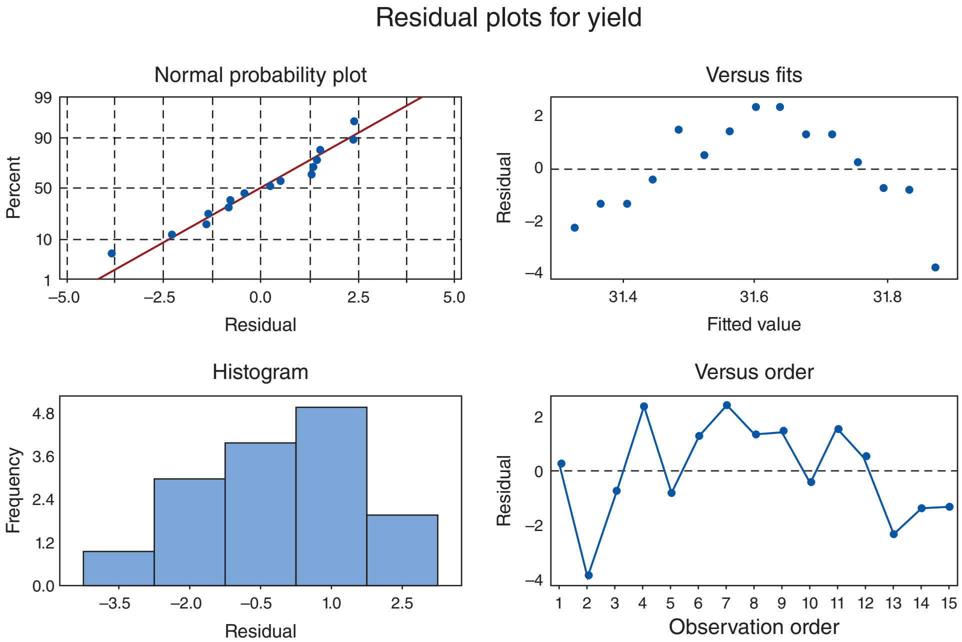


Figure 15.7.4 MINITAB residual plots for the data in Example 15.7.1.

In the above results, we note that the R-square value is only 0.95% and that indicates a lack of fit. However, the R-square quantity can produce meaningless results in the presence of nonlinear data and one should not solely make inferences depending on this quantity. Moreover, the p-value for the test of hypothesis $\beta_1 = 0$ is 0.73. This implies that there is no significant linear relationship between the yield and phosphate. This inference seem reasonable since the data show no considerable deviation from the normality as shown in the normal probability plot and histogram and the observations seem independent as the plot of residuals versus observation orders does not show any unusual pattern that would violate the assumption that the ε_i's are independent. The residuals show no extreme observation present in the data apart from the second observation being comparatively smaller than the rest. However, the plot of residual versus fitted values clearly suggests that the model is missing possibly a quadratic term of the independent variable.

USING R

The following R-code can be used to obtain the necessary similar outputs in R.

```
Y = c(32,28,31,34,31,33,34,33,33,31,33,32,29,30,30)
X = c(26,20,24,32,22,28,34,30,36,42,40,38,48,44,46)

#Fitting LSR model
model = lm(Y~X)

#ANOVA and Summary outputs
anova(model)
summary(model)
```

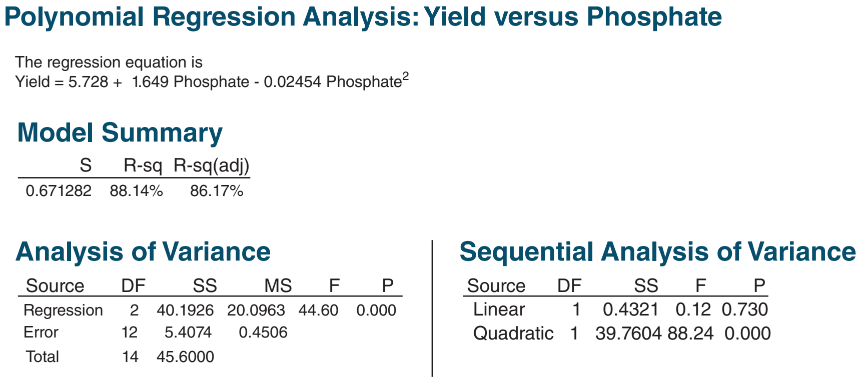
```
residuals =cbind(X,Y, round(cbind(model$fitted, model$res,rstandard(model)),2))
colnames(residuals) = c(“Phosphate”, “Yield”, “Fits”, “Residuals”, “StdRes.”)
residuals #This will produce residuals

#Residual plots
par(mfrow=c(2,2))
boxplot(model$res, ylab=“Residual”, main=“Boxplot of Residuals”, col=5)
qqnorm(model$res,pch=20, cex=2); qqline(model$res, col=2)
plot(model$fitted,model$res, xlab=“Fitted Value”, ylab=“Residual”,
     main =“Versus Fits”, col=4, pch=20, cex =2)
plot(c(1:length(X)),model$res, xlab=“Observation Order”, ylab=“Residual”,
     type=“b”, main =“Versus Order”, pch=20, cex =2, col =4)
abline(h=0, lty =2)
```

From the MINITAB and R outputs above, we have ample evidence to suggest that the inclusion in the model of quadratic trend may have provided a better fit to the data. We can then disregard the linear regression analysis as unsuitable and proceed to fit a second-order regression model (see Chapter 16).

MINITAB

To fit a second-order regression model for the data in Example 15.7.1 using MINITAB we proceed as follows.

From the Menu bar select **Stat** > **Regression** > **Fitted line plot** and then check the **Quadratic** option and other options to have the new residual plots. Then click **OK**. The following results appear in the Session window.



```
Polynomial Regression Analysis: Yield versus Phosphate

The regression equation is
Yield = 5.728 +  1.649 Phosphate - 0.02454 Phosphate^2

Model Summary
       S    R-sq  R-sq(adj)
0.671282  88.14%     86.17%

Analysis of Variance                              | Sequential Analysis of Variance
Source      DF       SS       MS      F      P    | Source     DF       SS      F      P
Regression   2  40.1926  20.0963  44.60  0.000    | Linear      1   0.4321   0.12  0.730
Error       12   5.4074   0.4506                  | Quadratic   1  39.7604  88.24  0.000
Total       14  45.6000                           |
```

Additionally MINITAB produces Figures 15.7.5 and 15.7.6. All these outputs produced by the least-squares procedure that fits the model $\eta = E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$ to the data (see also Chapter 16).

From this analysis, we observe that the value of R-square at 88.1% is quite high. Moreover, the p-value for testing the coefficient of the quadratic term is zero. This means we reject the null hypothesis, $H_0 : \beta_2 = 0$ in favor of the alternative hypothesis $H_1 : \beta_2 \neq 0$. In the ANOVA table the p-value for the regression is zero, which suggests that the quadratic model (Yield = 5.728 + 1.649 Phosphate − 0.02454 Phosphate^2) explains the relationship between the phosphate and the yield of soybean quite well.

Figure 15.7.6 shows the fitted regression model including its 95% confidence and prediction intervals. It is clear that the model is fitted to the data quite well and the residual plots in Figure 15.7.5 show no abnormalities about the quadratic model.

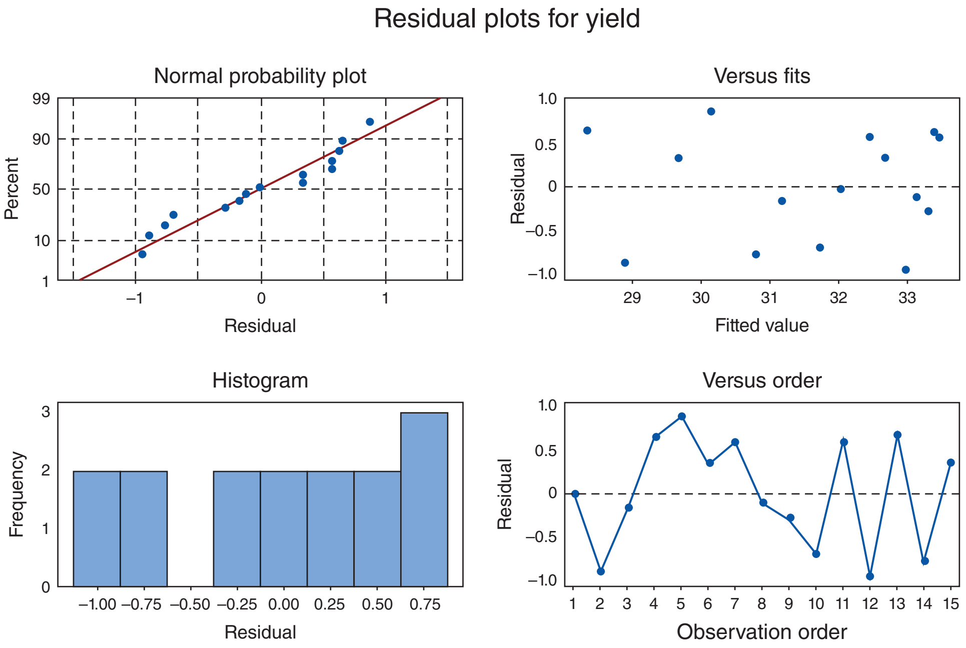


Figure 15.7.5 MINITAB plots for residual plots for yield and box plot of residuals when the quadratic model is fitted.

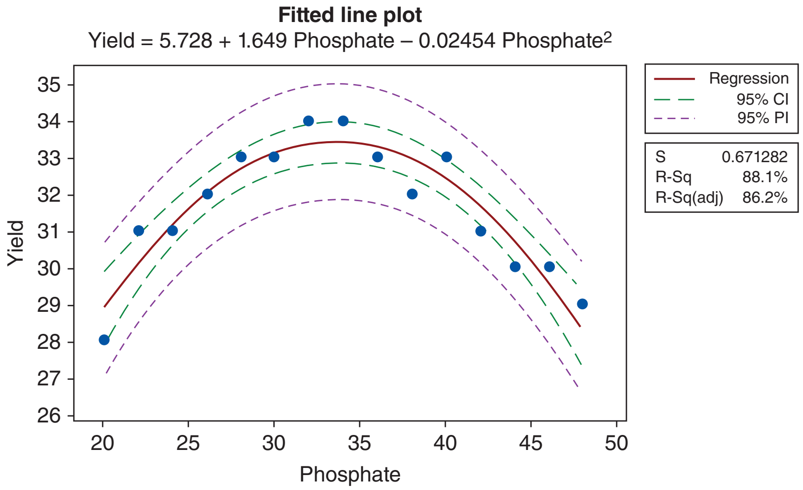


Figure 15.7.6 MINITAB plot for fitted quadratic: yield versus phosphate.

USING R

Use the following R-code to conduct the polynomial regression model for this example. Note here that to include the quadratic term, we use the index function ‘I’ in lm(Y∼ X+I(X^2)).

```
#Polynomial regression model
model = lm(Y~ X+I(X ˆ 2))
summary(model) # This will produce a summary table
anova(model) # This will produce an ANOVA table

residuals = cbind(Y,X,model$res/sqrt(1.864), model$res,model$fitted)
colnames(residuals) = c(“Yield”, “Phosphate”,“StdRes.”, “Residuals”, “Fits”)
residuals #This will produce residuals

#Residual plots
par(mfrow=c(2,2))
boxplot(model$res, ylab=“Residual”, main=“Boxplot of Residuals”, col=5)
qqnorm(model$res, pch=20, cex=2); qqline(model$res, col=2)
plot(model$fitted, model$res, xlab=“Fitted Value”, ylab=“Residual”,
main =“Versus Fits”, col=4, pch=20, cex =2)
plot(c(1:length(X)), model$res, xlab=“Observation Order”, ylab=“Residual”,
     type=“b”, main =“Versus Order”, pch=20, cex =2, col =4)
abline(h=0, lty =2)
```

The results that would appear are virtually identical to the results obtained using MINITAB. That is, based on the data, the quadratic model explains the relationship between phosphate and yield better than the first-order model.

PRACTICE PROBLEMS FOR SECTION 15.7

In the following problems reference is made to the problems for Section 15.2.

1. Refer to Problem 1.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.

2. Refer to Problem 2.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
3. Refer to Problem 3.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
4. Refer to Problem 4.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
5. Refer to Problem 5.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
6. Refer to Problem 6.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
7. Refer to Problem 7.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
8. Refer to Problem 8.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
9. Refer to Problem 9.
   (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
   (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.

10. Refer to Problem 10.
    (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
    (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
11. Refer to Problem 11.
    (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
    (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.
12. Refer to Problem 12.
    (a) Construct a normal probability plot of the residuals obtained from the least-squares fit. Does this plot indicate that the normality assumption is valid?
    (b) Plot the residuals versus fitted values, and order of the observations, and interpret these plots.

15.8 TRANSFORMATIONS

To begin our discussion in this section, a definition is in order.

Definition 15.8.1 Suppose that a random variable, Y, a response variable that is thought to be dependent on k independent (predictor) variables, $X_1, X_2, \dots, X_k$, so that $E(Y) = f(X_1, X_2, \dots, X_k)$. We then say that the model

$$E(Y) = \sum_{i=0}^{k} \beta_i X_i \tag{15.8.1}$$

is a *linear model*, where $X_0 = 1$. The quantities $\beta_0, \beta_1, \dots, \beta_k$ are parameters, usually unknown, and we note that they enter into the model *linearly.* A model containing quadratic or higher order terms of these parameters, *regardless of the degree of the predictor variables*, is called nonlinear.

Some examples of linear and nonlinear models are

$E(Y) = \beta_0 + \beta_1 X$	First-order linear model
$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$	Quadratic or second-order linear model
$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$	kth-Order linear model
$E(Y) = (\beta_0 + \beta_1 X)^{-1}$	Nonlinear model
$E(Y) = \alpha X^\beta$	Nonlinear model
$E(Y) = \alpha e^{\beta X}$	Nonlinear model

So far most of our discussion has centered on the linear regression model given in (15.2.4), which contains one response and one predictor variable, and these are related linearly. That model also assumes that ε_i, $i = 1, 2, \ldots, n$, are independently normally distributed with mean 0 and constant variance σ^2. In practice, however, we rarely meet all of these conditions, or we may meet these conditions yet still find the model inadequate. For instance, in Example 15.7.1, the model (15.2.4) is not adequate since the residual analysis indicates that the model may need an additional quadratic term. Indeed, after adding a quadratic term, we found previously, that the quadratic model was quite adequate. So that the new model is quadratic, but linear in parameters and satisfies all other assumptions. In many other instances, however, either our model is linear and the assumptions of the model (15.2.4) are not met, or, our model is not linear, nor are the assumptions of the model (15.2.4) met. One remedy for dealing with these problems is by making some transformations on the response variable, the predictor variable, or both.

Normality and stabilization of variance often go hand in hand. To stabilize the variance, transformations, such as the log, square root, or reciprocal of the response variables, are usually used. If the normality, common variance, and independence assumptions are fairly satisfactory, then, for linearization, transformations of the predictor variable such as X^2, log X, $\sqrt{X}$ or X^{-1} may work quite well. However, several transformations should be tried on the response variable, predictor variable, or both, and the residual analysis done to find which transformation works best. Sometimes one transformation may need to be used on the response variable and another on the predictor variable. Other transformations, for example, to linearize the nonlinear models mentioned above that may work well, are

Model	Transformation
$E(Y) = (\beta_0 + \beta_1 X)^{-1}$	Reciprocal transformation on response variable
$E(Y) = \alpha X^\beta$	Log transformation on both the response and predictor variable
$E(Y) = \alpha e^{\beta x}$	Natural log transformation on the response variable

Sometimes such transformations may also address other problems such as normalization and stabilization of variance. Sometimes many transformations are tried in order to find the one that works for linearization, normalization, and stabilization of the variance. Box and Cox (1964) provided a family of power transformations and a procedure, using the maximum likelihood method under the normality assumption, to determine an appropriate transformation. The essence of their procedure is as follows. In model (15.2.4), we would consider the response variable to be Y^λ and rewrite (15.2.4) as

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{15.8.2}$$

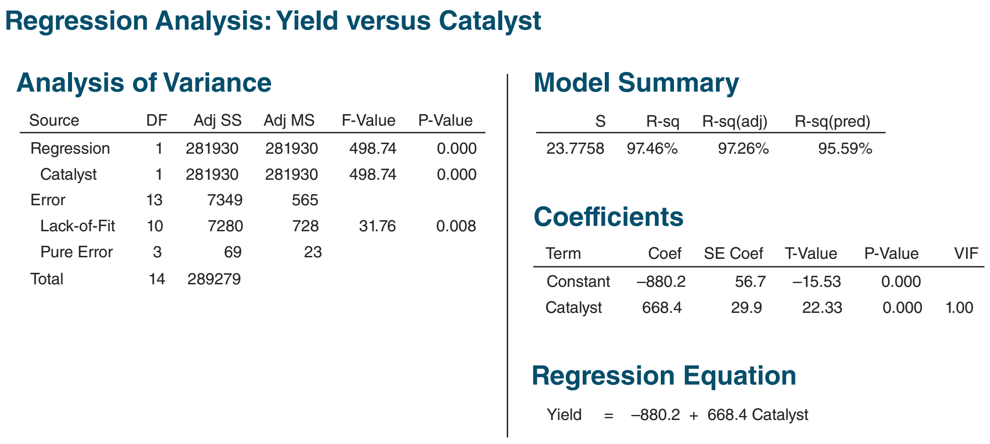
Now, using the method of maximum likelihood, we would determine the estimates $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\lambda}$ for β_0, β_1, and λ, respectively, and then use the transformation $Y^{\hat{\lambda}}$ on the response variable.

Table 15.8.1 Data for the experiment in Example.

Catalyst	1.91	1.88	2.16	1.83	1.88	1.67	2.21	1.86
Yield	382.84	355.36	564.14	325.83	361.86	256.02	592.70	349.52
ln(yield)	5.95	5.87	6.34	5.79	5.89	5.55	6.38	5.86
Catalyst	1.67	2.33	1.62	1.95	1.70	1.67	1.90	
Yield	247.42	727.58	235.15	385.05	265.60	247.72	375.45	
ln(yield)	5.51	6.59	5.46	5.95	5.58	5.51	5.93	

Example 15.8.1 (Chemical yield versus amount of catalyst) *It is believed that the yield of a chemical depends on the amount of catalyst used. An experiment using 15 randomly selected amounts of catalyst is run to produce the data recorded in Table 15.8.1. We include the natural log of the yield for reasons to be discussed later in the solution. Fit an appropriate regression model to the data in Table 15.8.1.*

Solution: We first start with the simplest regression model (15.2.4) using catalyst as the predictor variable and yield as the response variable. MINITAB produces the following output for the regression analysis.



```
Regression Analysis: Yield versus Catalyst

Analysis of Variance
Source          DF   Adj SS   Adj MS   F-Value   P-Value
Regression       1   281930   281930    498.74     0.000
  Catalyst       1   281930   281930    498.74     0.000
Error           13     7349      565
  Lack-of-Fit   10     7280      728     31.76     0.008
  Pure Error     3       69       23
Total           14   289279

Model Summary
      S    R-sq   R-sq(adj)   R-sq(pred)
23.7758  97.46%      97.26%       95.59%

Coefficients
Term        Coef   SE Coef   T-Value   P-Value    VIF
Constant  –880.2      56.7    –15.53     0.000
Catalyst   668.4      29.9     22.33     0.000   1.00

Regression Equation
Yield  =  –880.2 + 668.4 Catalyst
```

From the MINITAB output, we note that the value of the coefficient of determination R^2 is very high (97.46%) and the p-values for both hypotheses, $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$ and $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, are zero. Hence, we reject the null hypotheses in favor of the alternative hypotheses. In other words, both β_0 and β_1 are significantly different from zero, and the temptation is to conclude that the simple linear model is an adequate fit. However, if we examine the results more closely, suspicions about our conclusions arise. First, we note that the standard errors of both $\hat{\beta}_0$ and $\hat{\beta}_1$ are quite high, as is S, the estimate of σ. Furthermore, the plot of residual versus fitted value in Figure 15.8.1 presents some anomalies, whereas normal probability, histogram, and residual versus the ordered observations plots are satisfactory, that is the conditions of normality and independence are valid. Thus, despite the high value of R^2 and very low p-values, our conclusion of linearity does not seem satisfactory.

A careful investigation of the fitted regression line in the Figure 15.8.2a indicates that a slight exponential type growth in yield with respect to the catalyst. This indicates that the natural log transformation on the response variable may produce a better fit. One can experiment with various alternative transformations on the response variable to obtain even better results. However, the resulting fitted model seen in Figure 15.8.2, indicates that the suggested transformation adequately linearizes the relationship between the variables. Here, we present the MINITAB output of the regression analysis of ln(Yield) versus catalyst.

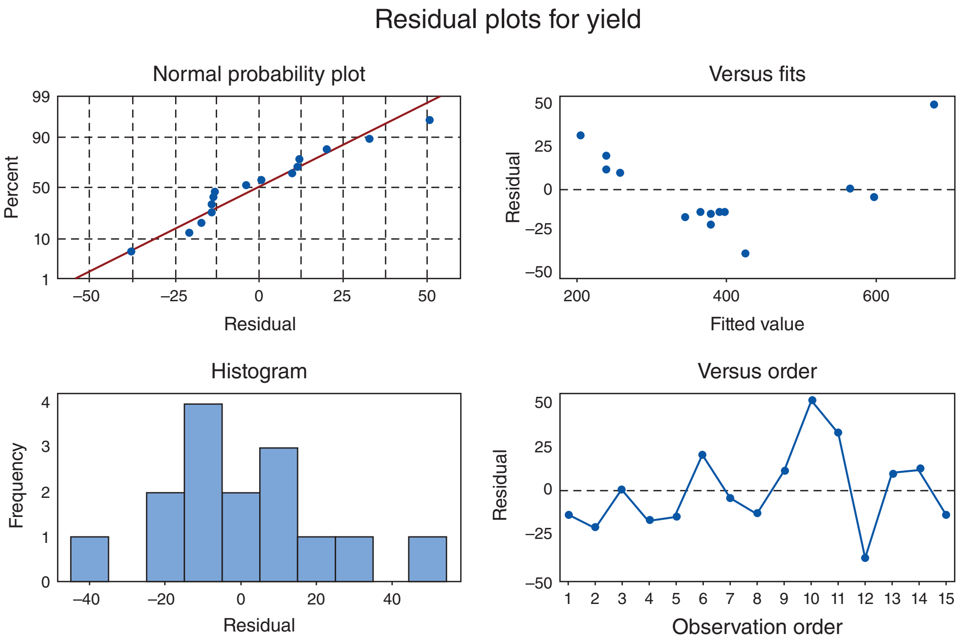


Figure 15.8.1 MINITAB residual plots: regression analysis of yield versus catalyst.

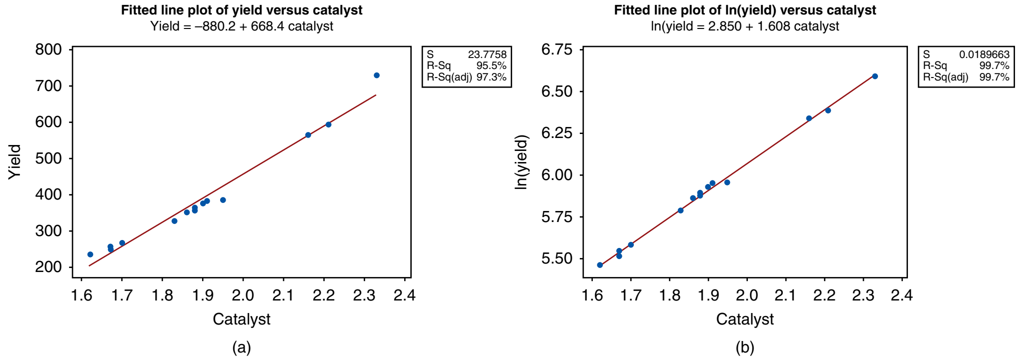


Figure 15.8.2 Fitted regression lines for (a) yield versus catalyst and (b) ln(yield) versus catalyst.

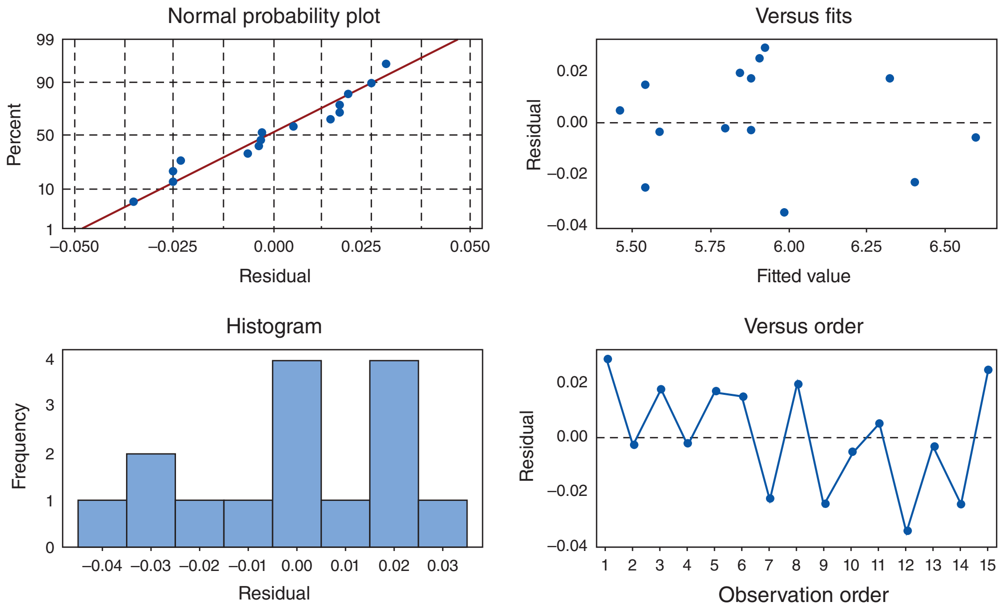
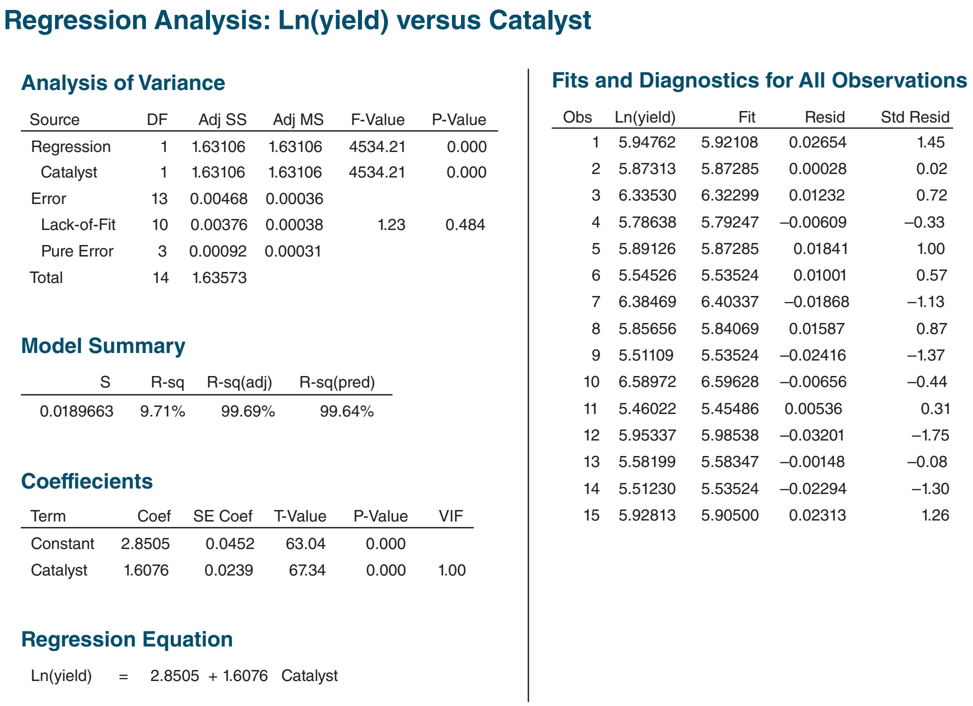


Figure 15.8.3 MINITAB residual plots for the regression analysis of ln(yield) versus catalyst.



Regression Analysis: Ln(yield) versus Catalyst

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1.63106	1.63106	4534.21	0.000
Catalyst	1	1.63106	1.63106	4534.21	0.000
Error	13	0.00468	0.00036		
Lack-of-Fit	10	0.00376	0.00038	1.23	0.484
Pure Error	3	0.00092	0.00031		
Total	14	1.63573			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0189663	9.71%	99.69%	99.64%

Coeffiecients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.8505	0.0452	63.04	0.000	
Catalyst	1.6076	0.0239	67.34	0.000	1.00

Regression Equation

Ln(yield) = 2.8505 + 1.6076 Catalyst

Fits and Diagnostics for All Observations

Obs	Ln(yield)	Fit	Resid	Std Resid
1	5.94762	5.92108	0.02654	1.45
2	5.87313	5.87285	0.00028	0.02
3	6.33530	6.32299	0.01232	0.72
4	5.78638	5.79247	−0.00609	−0.33
5	5.89126	5.87285	0.01841	1.00
6	5.54526	5.53524	0.01001	0.57
7	6.38469	6.40337	−0.01868	−1.13
8	5.85656	5.84069	0.01587	0.87
9	5.51109	5.53524	−0.02416	−1.37
10	6.58972	6.59628	−0.00656	−0.44
11	5.46022	5.45486	0.00536	0.31
12	5.95337	5.98538	−0.03201	−1.75
13	5.58199	5.58347	−0.00148	−0.08
14	5.51230	5.53524	−0.02294	−1.30
15	5.92813	5.90500	0.02313	1.26

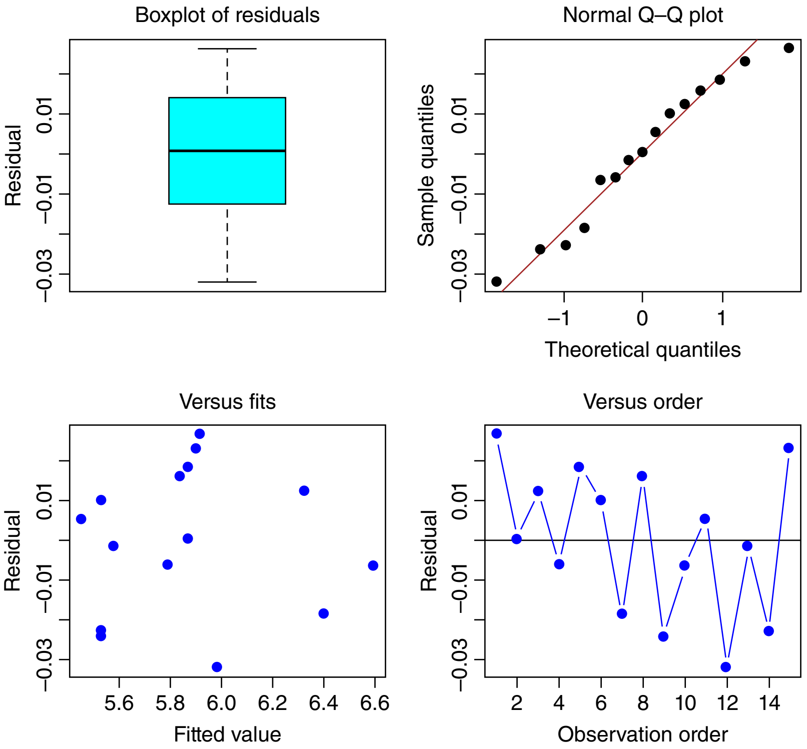


Figure 15.8.4 R residual plots for the regression analysis of ln(Yield) versus catalyst.

The standard errors of both $\hat{\beta}_0$ and $\hat{\beta}_1$ are very small, and so is S, the estimate of σ. Also the value of R^2 (99.7%) is very high and the p-values for both hypotheses, $H_0 : \beta_1 = 0$ versus $H_1 : \beta_0 \neq 0$ and $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ are zero. Moreover, the residual plots in Figure 15.8.3 do not present any particular anomalies. Thus, the *ln* transformation on the response variable gives a better fit.

Solution USING R

To perform a regression analysis on ln(Yield), we can use the following R-code. *Note*: lm(log(Yield) ∼ Catalyst) is directly used to regress ln(Yield) against the catalyst. We find that this output is identical to that obtained in MINITAB, and we may then draw the same conclusions as well.

```
Catalyst = c(1.91,1.88,2.16,1.83,1.88,1.67,2.21,1.86,1.67,2.33,1.62,1.95,1.70,
   1.67,1.90)
Yield = c(382.84,355.36,564.14,325.83,361.86,256.02,592.70,349.52,247.42,727.58,
   235.15,385.05,265.60,247.72,375.45)

#Fitting LSR model with ln(Yield)
model = lm(log(Yield) ∼ Catalyst)
summary(model) # This will produce a summary table
```

```
anova(model) # This will produce an ANOVA table

residuals = cbind(Catalyst, log(Yield), round(cbind(model$fitted, model$res, rstan-
dard(model)),4))
colnames(residuals) = c(“Catalyst”, “log(Yield)”, “Fits”, “Residuals”, “StdRes.”)

#Residual plots
par(mfrow=c(2,2))
boxplot(model$res, ylab=“Residual”, main=“Boxplot of Residuals”, col=5)
qqnorm(model$res,pch=20, cex=2); qqline(model$res, col=2)
plot(model$fitted,model$res, xlab=“Fitted Value”, ylab=“Residual”,
     main =“Versus Fits”, col=4, pch=20, cex =2)
plot(c(1:length(Catalyst)),model$res, xlab=“Observation Order”, ylab=“Residual”,
     type=“b”,main =“Versus Order”, pch=20, cex =2, col =4)
abline(h=0, lty =2)
```

The residual plots for the regression analysis of ln(Yield) versus catalyst using R are shown in Figure 15.8.4. These graphs convey the same information provide by the MINITAB residual plots shown in Figure 15.8.3. Both MINITAB and R plots do not present any particular anomalies. Thus, the previous conclusion that the *ln* transformation on the response variable gives a better fit is further validated.

PRACTICE PROBLEMS FOR SECTION 15.8

In the following problems, reference is made to the problems in Section 15.2

1. Refer to Problem 4.
   (a) Fit a simple linear regression model using log Y as the response variable.
   (b) Using (a), record the least-squares estimates of the regression coefficients.
   (c) Test for significance of the regression coefficient. That is, test the hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ for $i = 0$ and $i = 1$. Use $\alpha = 0.05$.
2. Refer to Problem 10.
   (a) Fit a simple linear regression model using $\sqrt{Y}$ as the response variable.
   (b) Using (a), record the least-squares estimates of the regression coefficients.
   (c) Test for significance of the regression coefficients. That is, test the hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ for $i = 0$ and $i = 1$. Use $\alpha = 0.05$.
3. Refer to Problem 4.
   (a) Fit a simple linear regression model using log Y and log X as the response variable and the predictor variable, respectively.
   (b) Using (a), record the least-squares estimates of the regression coefficients.
   (c) Test for significance of the regression coefficients. That is, test the hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ for $i = 0$ and $i = 1$. Use $\alpha = 0.05$.

15.9 INFERENCE ABOUT ρ

Quite often we wish to find confidence intervals and make tests of hypotheses concerning the parameter ρ, the population correlation coefficient. Suppose that $(X_i, Y_i), i = 1, \ldots, n$, are n independent observations on the bivariate random variable (X, Y), where (X, Y) has the bivariate normal distribution. The sampling distribution of the *sample correlation coefficient* r, where r is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, -1 \leq r \leq 1$$

which can be written as

$$\boxed{r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \qquad (15.9.1)}$$

is quite complicated for general values of the population correlation coefficient $\rho, -1 \leq \rho \leq 1$. We emphasize that both components of each observation are random, and their joint distribution is bivariate normal. This is unlike the previous cases discussed in this chapter where X_i are assumed to be fixed constants chosen before the Y_i are observed at X_i.

Now, for the case when $\rho = 0$, it is known that the distribution of r is such that the quantity

$$\boxed{\tfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (15.9.2)}$$

has the Student t-distribution with $(n-2)$ degrees of freedom. Hence, for example, to test the hypothesis

$$H_0 : \rho = 0 \quad \text{versus} \quad H_1 : \rho \neq 0$$

at the α level of significance, we use (15.9.2) as our test statistic and reject H_0 in favor of H_1 if the observed value of

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{n-2;\alpha/2} \qquad (15.9.3)$$

Example 15.9.1 (Carbon contents in ball clay) *Bennett and Franklin (1954) discuss the analysis for carbon content of 36 specimens of ball clay by two different methods. Let* X *= carbon determined by combustion and* Y *= carbon determined by "rational analysis,"* *where both* X *and* Y *are measured in percent. It turns out that the values of* (X_i, Y_i), $i = 1, \ldots, 36$ *for the 36 specimens gave the following results:*

$$n = 36, \quad \sum X_i = 69.25, \quad \sum Y_i = 102.71, \quad \sum X_i^2 = 354.0245, \quad \sum Y_i^2 = 826.6842,$$
$$\sum X_i Y_i = 510.8425$$

From these we find that

$$S_{XX} = \sum (X_i - \bar{X})^2 = 220.8144, \quad S_{YY} = \sum (Y_i - \bar{Y})^2 = 533.6469$$

and

$$\begin{aligned} S_{XY} &= \textstyle\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i\, Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)/n \\ &= 510.8425 - (69.25)(102.71)/36 \\ &= 510.8425 - 197.5741 = 313.2684 \end{aligned}$$

Hence

$$r = \frac{313.2684}{\sqrt{(220.8144)(533.6469)}} = 0.91259$$

Thus, the observed value of the test statistic in (15.9.2) is

$$\frac{0.91259}{\sqrt{1 - (0.91259)^2}}\sqrt{34} = 13.014$$

a highly significant result at either the 5% or 1% level of significance, and so we reject the null hypothesis $H_0 : \rho = 0$.

Thus, it is highly unlikely that the population correlation coefficient, ρ, could be zero, so we may wish to instead find a confidence interval for ρ. Now, it can be shown that for large n, if $(X_1, Y_1), \ldots, (X_n, Y_n)$ can be regarded as a random sample from a *bivariate normal distribution* having correlation coefficient ρ, then the quantity

$$\left(\frac{1}{2}\right) \log_e \frac{1+r}{1-r} = \tanh^{-1} r \tag{15.9.4}$$

is approximately normally distributed with mean

$$\tau = \left(\frac{1}{2}\right) \log_e \frac{1+\rho}{1-\rho} = \tanh^{-1} \rho \tag{15.9.5}$$

and variance $1/(n-3)$. By using (15.9.4) and (15.9.5), we see that an approximate 100 α% confidence interval for

$$\tau = \left(\frac{1}{2}\right) \log_e \frac{1+\rho}{1-\rho}$$

is given by

$$\left(\left(\frac{1}{2}\right) \log_e \frac{1+r}{1-r} \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}}\right) \tag{15.9.6}$$

Consulting normal tables, we have that for the Example 15.9.1, with $1 - \alpha = 0.95$, and substituting the values $r = 0.9126$, $n = 36$, the confidence interval for $(\frac{1}{2})\log_e[(1 + \rho)/(1 - \rho)]$ is

$$\left(1.543 \pm 1.96\frac{1}{\sqrt{33}}\right) = (1.202, 1.884) \tag{15.9.7}$$

Now since $\tau = (\frac{1}{2})\log_e((1 + \rho)/(1 - \rho))$ is increasing function in ρ, using (15.9.7) it can easily be shown that a 95% confidence interval for ρ is $(0.83, 0.95)$, as follows. Since τ is increasing function in ρ we have that

$$\begin{aligned} P\left(\tau_l \le \left(\frac{1}{2}\right)\log_e\frac{1+\rho}{1-\rho} \le \tau_u\right) &= P\left(\frac{e^{2\tau_l} - 1}{e^{2\tau_l} + 1} \le \rho \le \frac{e^{2\tau_u} - 1}{e^{2\tau u} + 1}\right) \\ &= P\left(\frac{e^{2(1.202)} - 1}{e^{2(1.202)} + 1} \le \rho \le \frac{e^{2(1.884)} - 1}{e^{2(1.884)} + 1}\right) \end{aligned}$$

or

$$P(0.83 \le \rho \le 0.95) = 0.95$$

that is, $(0.83, 0.95)$ is a 95% confidence interval for ρ.

PRACTICE PROBLEMS FOR SECTION 15.9

In the following problems, reference is made to the problems for Section 15.2.

1. Refer to Problem 1. Test at the 5% level of significance the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.
2. Refer to Problem 3. Test at the 5% level of significance the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.
3. Refer to Problem 4. Test at the 5% level of significance the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.
4. Refer to Problem 7. Test at the 5% level of significance the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.
5. Refer to Problem 10. Test at the 5% level of significance the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.
6. Refer to Problem 11. Test at the 5% level of significance the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$.

15.10 A CASE STUDY

Case Study (*Load cell calibration*)[1] The data collected in this calibration experiment consisted of a known load applied to a load cell and the corresponding deflection of the cell from its nominal position. Forty measurements were made over a range of loads from 150,000 to 3,000,000 units. The data were collected in two sets in order of increasing load. The systematic run order makes it difficult to determine whether there was any drift in the load cell or measuring equipment over time. Assuming there is no drift, however, the

[1] Source: NIST and SEMATECH (2003).

experiment should provide a good description of the relationship between the load applied to the cell and its response. The resulting data are given below

Deflection	Load	Deflection	Load
0.11019	150,000	0.11052	150,000
0.21956	300,000	0.22018	300,000
0.32949	450,000	0.32939	450,000
0.43899	600,000	0.43886	600,000
0.54803	750,000	0.54798	750,000
0.65694	900,000	0.65739	900,000
0.76562	1,050,000	0.76596	1,050,000
0.87487	1,200,000	0.87474	1,200,000
0.98292	1,350,000	0.98300	1,350,000
1.09146	1,500,000	1.0915	1,500,000
1.20001	1,650,000	1.20004	1,650,000
1.30822	1,800,000	1.30818	1,800,000
1.41599	1,950,000	1.41613	1,950,000
1.52399	2,100,000	1.52408	2,100,000
1.63194	2,250,000	1.63159	2,250,000
1.73947	2,400,000	1.73965	2,400,000
1.84646	2,550,000	1.84696	2,550,000
1.95392	2,700,000	1.95445	2,700,000
2.06128	2,850,000	2.06177	2,850,000
2.16844	3,000,000	2.16829	3,000,000

(a) In this problem, indicate which variable is independent and which is dependent.
(b) Fit a first-order linear regression model to the load and deflection data.
(c) Perform the residual analysis and check the adequacy of the model.
(d) If in (b) the model is not adequate, then fit a second-order linear regression model to the load and deflection data.
(e) Perform the residual analysis and again check the adequacy of the model.
(f) If in part (d) the model is adequate, then use it for predicting future observations. Develop a 95% confidence interval for the individual and the average value of the dependent variable.

15.11 USING JMP

This section is not included in the book but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Suppose that it has been assumed on prior theoretical grounds that $E(Y|X) = \eta = \beta_1 X$. To gain knowledge of β_1, it is decided to experiment by setting the independent

variable X to each of the values $x_1, \ldots, x_n$ and to observe the resulting Y's, say $Y_1, \ldots, Y_n$. Determine by the method of least squares an estimate b_1 of β_1 and hence an estimator of this regression line of Y on X (which passes through the origin). Find the expectation and variance of b_1, assuming that the Y_i are independent, with expectation $\beta_1 x_i$ and variance σ^2. If the Y_i are, in addition, normally distributed, state how you would determine confidence intervals for β_1 and the true response η_0 when $X = x_0$.

2. It is decided to measure the resistance of sheets of a certain metal at temperatures X of 100, 200, 300, 400, and 500 K. The resistances Y are found to be 4.7, 7.4, 12.4, 16.5, and 19.8, respectively. If the regression of Y on X is assumed to be linear, state the normal equations for the parameters in the linear model, and solve.

3. An experiment is planned in which three observations will be taken at each of four temperatures, 30°, 50°, 70°, and 90°. When the experiment is actually done, the results obtained are those given below. Find, assuming $\eta = E(Y|X) = \beta_0 + \beta_1 X$, the least-squares estimates of β_0 and β_1, and hence the least-squares estimate of η. Construct 95% confidence intervals for β_0, β_1, and $\eta_0 = \beta_0 + \beta_1 x_0$.

X (temperature)	30	30	30	50	50	50	70	70	70	90	90	90
Y (response)	40	45	31	34	28	35	21	29	25	16	21	23

4. A chemical engineer wants to fit a straight-line to the data found observing the tensile strength, Y, of 10 test pieces of plastic that have undergone baking (at a uniform temperature) for X minutes, where 10 values of X were preselected. The data (in coded units) is given below. Repeat the instructions of Problem 3 for this set of data.

X	23	35	45	65	75	95	105	125	155	185
Y	2	9.8	9.2	26.2	17.1	24.8	43	55.3	38.4	63.3

5. An investigation of the (assumed) linear relationship between the load X on a spring and the subsequent length of the spring Y has been carried out with the results given below. Repeat the instructions of Problem 3 for this set of data.

X	5	10	15	20	25	30
Y	7.25	8.12	8.95	9.90	10.9	11.8

6. A study was made on the effect of pressure X on the yield Y of paint made by a certain chemical process. The results (in coded units) are given below. Repeat the instructions of Problem 3. In the notation of Problem 3 above, let X_0 take the values −5, −3, −1, 1, 3, and 5 and draw the confidence band for the estimated regression equation.

X	−5	−4	−3	−2	−1	0	1	2	3	4	5
Y	6.5	10.3	9.7	12.1	15.7	13.7	14.2	18.0	19.7	18.8	23.4

7. The relationship (assumed linear) between the yield of bourbon Y and aging time X was studied by observing yields Y_i from batches that have been allowed to age $X_i = 2i$ years, $i = 1, 2, \ldots, 6$. The results are given below. Repeat the instructions of Problem 6 for this set of data, but use $X_0 = 2, 4, 6, 8, 10,$ and 12 for computing the confidence band for the estimated regression equation.

X	2	4	6	8	10	12
Y	3.0	3.4	4.0	4.5	4.4	5.0

8. The moisture X of the wet mix of a product is considered to have an effect on the density Y of the finished product. The moisture of the mix was controlled and the finished product densities were as shown below. Repeat the instructions of Problem 6 for this set of data, choosing convenient values of X_0 for drawing the confidence band.

X	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	7.6	9.5	9.3	10.3	11.1	12.1	13.3	12.7	13.0	13.8	14.6	14.8	14.7

9. The effect of temperature (X in Kelvin) on the color (Y, coded units) of a product was investigated and the results obtained are given below. Construct an analysis of variance table for these data and then use it to test the hypothesis $H_0 : \beta_1 = 0$. If the test rejects H_0, then fit these data to the model $E(Y|X) = \beta_0 + \beta_1 X$. What is the estimate of $E(Y|X)$ when $X = 145$? Find a 95% confidence interval for $E(Y|145)$.

X	100	110	120	140	150	170	180	200	230	260
Y	3.5	7.4	7.1	15.6	11.1	14.9	23.5	27.1	22.1	32.9

10. Specimens of blood from 10 different animals were analyzed for blood count, say, Y (in units of 100) and packed cell volume count X (in mm^3) with results as given below. Assuming normality, test the hypothesis that the true correlation coefficient ρ between blood count and volume count is zero. If the test is rejected, use the method of Section 15.19 to find a 95% confidence interval for ρ.

Animal #	X	Y
1	45	6.5
2	42	6.3
3	56	9.5
4	48	7.5
5	42	7.0
6	35	5.9
7	58	9.5
8	40	6.2
9	39	6.6
10	50	8.7

11. Repeat the instructions of Problem 10 for the data given below. The data set was obtained by measuring the tensile strength (Y) in 1000 psi and the Brinell

hardness (X) of each of 15 specimens of cold-drawn copper (data from Bowker and Lieberman, 1959).

Specimen #	X	Y
1	104.2	39.8
2	106.1	40.4
3	105.6	39.9
4	106.3	40.8
5	101.7	33.7
6	104.4	39.5
7	102.0	33.0
8	103.8	37.0
9	104.0	37.6
10	101.5	33.2
11	101.9	33.9
12	100.6	29.9
13	104.9	39.5
14	106.2	40.6
15	103.1	35.1

12. Repeat the instructions of Problem 10 for the data given below. These data (Bullis and Alderton) were obtained by examining the alpha resin content of six different specimens of hops by taking colorimeter readings X and by direct determination of the concentration (Y, in mg per 100 ml).

Specimen #	X	Y
1	8	0.12
2	50	0.71
3	81	1.09
4	102	1.38
5	140	1.95
6	181	2.05

13. Test pieces of boiler plate undergo tests at various times during a production process. The measurements made are X, the force applied in tons per square inch at the time of removal from the process, and the resulting elongation Y of the test piece. For the results given below on 10 test pieces, repeat the instructions of Problem 10.

Test piece #	X	Y
1	1.33	27
2	2.68	50
3	3.57	67
4	4.46	83
5	5.35	101
6	6.24	117
7	7.14	134
8	8.93	154
9	9.82	188
10	10.70	206

14. Experiments have shown that when a clean tungsten surface is heated by laser irradiation using a focused ruby laser, the rate of evaporation of tungsten from the surface is similar to that obtained by more conventional surface-heating methods. The following experimental data relates the observed temperature change as a function of laser amplitude. Fit a first-order polynomial to these data.

Laser pulse amplitude	0.7	0.9	1.1	1.2	1.5	1.7	1.9	2.0	2.1
Surface temperature (°C)	630	740	700	745	1120	1205	1510	1530	1520

15. In an investigation of pure copper bars of a small diameter, the following shear stress and shear strain data were collected.

Shear strain (%)	8.8	9.3	10.4	11.2	12.3	13.0	13.8	14.4	15.8	16.7	17.3
Shear stress (1000 psi)	11.8	11.9	11.4	11.6	12.3	12.7	13.3	13.7	13.8	14.4	14.5

(a) Fit a straight-line to the shear strain data as the independent variable and the shear stress data as the response variable.
(b) Refit the line with the roles of shear strain and shear stress reversed.
(c) Why do the two fitted models disagree?

16. In a study of internal combustion engines, the data given were observed with Y (in units of BTU/lb), the net work provided by a cylinder, as a function of the fuel fraction X, where $0 < X \leq 1$.

Fuel fraction	0.15	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.0
Net work	120	165	204	238	296	373	410	462	520	580	600
BTU/lbs of air							403	455	518		
							420	464	525		
							408				

Fit a straight-line to the data and test the hypothesis that the true slope is zero. Use the 5% significance level.

17. Sulfur dioxide can be removed from flue gases at low temperatures (approximately 600°F) through the use of a dry absorbent, alkalized alumina. The absorbent, when spent, is later regenerated in a separate process with elemental sulfur produced as a by-product. One series of experiments yielded the data below, relating the removal of sulfur dioxide as a function of the height of the absorber tower, under fixed operating conditions (the sulfur dioxide removed is a response or dependent variable and absorbent height is an independent or predictor variable):

Height of absorber (ft)	4.5	8.0	12.0	16.0	20.0	24.0	26.0
Removal of SO_2 (%)	7.2	21.2	28.3	33.0	42.0	54.0	63.8

(a) Fit a linear regression model to these data.
(b) What is the best estimate of the amount of sulfur dioxide to be removed in a tower 30 ft high?
(c) What is the estimate of the required height of the absorbent tower that is needed to remove 95% of the SO_2?

18. Estimate the acceleration, say τ, of a body disturbed from rest by a constant applied force. The approximate model is $V = \tau t$, where V is the velocity at time t.

t (s)	1	2	3	4	5	6	7	8
Velocity V (ft/s)	34.2	57.6	94.3	121.0	146.4	175.2	212.8	247

19. In Problem 18, find a 95% confidence interval for the slope parameter τ.

20. In Problem 13, find a 99% confidence interval for the observation Y and its expected value $E(Y|X = x)$, at $x = 5.50$ and $x = 8.20$.

21. In Problem 20, discuss whether it is reasonable to find a confidence interval for the observation Y or its expected value $E(Y|X)$ when $X = 11$.

22. In Problem 17, construct the ANOVA table and perform the F-test to learn whether the assumption that a linear regression model is the true model; do this at the 5% level of significance. State your conclusion.

23. In Problem 16, perform the residual analysis. Do you find any abnormalities? If you do find any abnormalities, then suggest some remedies so that a suitable model can be fitted.

24. The manager of a manufacturing company believes that experience is the most valuable variable in determining a worker's productivity. She collects data on productivity of 10 workers with known number of years of experience. The data collected are given below:

Experience, X	15	15	13	20	13	20	11	5	15	11
Productivity %, Y	116	128	116	130	106	120	124	111	129	105

(a) Fit a linear regression model to these data.
(b) Determine the coefficient of determination R^2.
(c) Determine the estimate of the error variance.
(d) Find a 95% confidence interval for the regression parameter β_1 and use this confidence interval to test the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. What is the level of significance for the test?

25. Refer to Problem 24, find the correlation coefficient between the X and Y variables. Perform a test, using the 5% level of significance, of the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho > 0$.

26. In Problem 24, use the log transformation on both the X and Y variables and fit the least-square line to the transformed data. Perform the residual analysis in this problem and in Problem 24 and compare the two fits. Give your conclusion about which fit is the better one. State the model that is being fitted in this problem.

27. A new cholesterol-lowering drug is being tested on eight randomly chosen patients. Since the appropriate dose of the drug is yet unknown, the chosen doses are varied and the resulting level of cholesterol changes in each patient are as given below:

Dose (mg), X	190	175	100	125	140	190	200	175
Cholesterol (mg/dL), Y	45	35	15	16	28	41	45	34

(a) Construct a scatter plot for these data and suggest an appropriate model that would provide a good fit to these data.
(b) Fit the suggested model and check if the suggested model is a good fit.
(c) Perform residual analysis and then indicate which, if any, of the assumptions of the model are being violated.

28. Refer to Problem 27. Assuming that a linear regression model is appropriate.
(a) Find a 95% prediction interval for Y when $X = 150\,\text{mg}$.
(b) Find a 95% confidence interval for $E(Y|X)$ when $X = 150\,\text{mg}$.
(c) Compare the confidence interval in (b) with the prediction interval in (a) and indicate which one is larger.

29. Consider the following data set.

X	6	8	11	15	17	20	25	30	33	40
Y	20	23	32	38	40	44	39	32	26	24

(a) Fit these data to the following models:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, 10$$
$$Y_i = \tau_0 + \tau_1 X_i + \tau_2 X_i^2 + \varepsilon_i, \quad i = 1, \ldots, 10$$

(b) Perform residual analysis for both fits and conclude which is a better one.

30. In Problem 29, after concluding which model is the "better fit," find a 99% prediction interval for Y and a 99% confidence interval for $E(Y|X)$ at $X = 22$ and 35, using the model with the better fit.

31. The following data are 28 observations Y on the yield of a certain by-product when certain temperatures X °C are used in a chemical process:

X	22	30	28	30	48	29	50	39	47	39	30	15	42	31
Y	72	91	69	81	71	85	79	77	81	75	73	62	70	65
X	28	10	3	12	19	33	27	4	27	36	46	12	17	8
Y	62	57	58	62	68	71	60	60	80	79	72	65	55	63

(a) Fit the model $E(Y|X) = \beta_0 + \beta_1 X$. Examine the residuals and comment on the linearity in X.
(b) Find a 95% confidence interval for β_0, and a 95% confidence interval for β_1.
(c) Find a 95% prediction interval for future observations at $X = 26$, 34, and 43.

32. The following data give 12 results of measuring the thickness Y of the silver film deposited when an amount X of a certain acid mix is used in a process. The values of X were preselected, and the corresponding values of Y are listed below in the order obtained:

X	4	6	2	5	7	6	3	8	5	3	1	5
Y	197	272	100	228	327	279	148	377	238	142	66	239

(a) Fit the model $E(Y|X) = \beta_0 + \beta_1 X$ to these data and do a residual analysis.
(b) Does the residual analysis in (a) support the assumption that the model is linear, that is, $E(Y|X) = \beta_0 + \beta_1 X$?
(c) If the answer in (b) is yes, carry out instructions in (b) of Problem 31.

33. Refer to Problem 3, Section 15.2.
(a) Determine the correlation coefficient between X and Y for the data in Problem 3.
(b) Test the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho > 0$. Use $\alpha = 0.05$ Find the observed level of significance (p-value). Give the practical interpretation of the p-value.

34. Refer to Problem 3, Section 15.2. Test the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho < 0$. Use $\alpha = 0.01$. Find the observed level of significance (p-value). Give the practical interpretation of the p-value.

35. Thirteen specimens of 90/10 Cu–Ni alloys, each with a specific iron content, were tested in a corrosion-wheel setup. The wheel was rotated in seawater at 30 ft/s for 60 days. The corrosion was measured in weight loss in milligrams/square decimeter/day, MDD. The data collected are given below (from Draper and Smith, 1981, used with permission):

X (Fe)	0.01	0.48	0.71	0.95	1.19	0.01	0.48
Y (loss in MDD)	127.6	124.0	110.8	103.9	101.5	130.1	122.0
X (Fe)	1.44	0.71	1.96	0.01	1.44	1.96	
Y (loss in MDD)	92.3	113.1	83.7	128.0	91.4	86.2	

(a) Fit a simple linear regression model to these data.
(b) Construct an ANOVA table for these data.
(c) Use the ANOVA you constructed in (b) to test the hypothesis (use $\alpha = 0.05$)

$$H_0 : \text{Fitted model is not appropriate} \quad \text{versus} \quad H_1 : \text{Fitted model is appropriate}$$

36. Refer to Problem 35.
(a) Estimate σ^2 and the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$.
(b) Find 95% confidence intervals for β_0 and β_1.

(c) Determine the sample correlation coefficient between X and Y.
(d) Test the hypothesis $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. Use $\alpha = 0.05$. Find the p-value.

37. Refer to Problem 35.
    (a) Fit a simple regression model using log Y as the dependent variable.
    (b) Estimate the value of Y at $X = 1.50$, using the model fitted in (a) and the model fitted in Problem 35. Compare the two results and comment on them.

38. Refer to Problem 35.
    (a) Determine all the residuals in Problem 35 and test if the normality assumption is satisfied.
    (b) Plot the residuals versus values of $\hat{Y}$. Comment on the assumption of constant variance of the Y_i.
    (c) Construct an appropriate graph to check the assumption of independence of the residuals.

Chapter 16

MULTIPLE LINEAR REGRESSION ANALYSIS

The focus of this chapter is the development of procedures to fit multiple linear regression models.

Topics Covered

- Multiple linear regression models
- Estimation of regression coefficients
- Estimation of regression coefficients using matrix notation
- Properties of the least-squares estimators
- Analysis of variance approach to regression analysis
- Discussion of inferences about the regression parameters
- Multiple linear regression models that use qualitative or categorical predictor variables
- Standardized regression coefficients, and multicollinearity and its consequences
- Building regression type prediction models
- Residual analysis
- Certain criteria for model selection
- Basic concepts of logistic regression

Learning Outcomes

After studying this chapter, the reader will be able to

- Use the least-squares method to estimate the regression coefficients in a multiple regression model and carry out hypothesis testing to determine which regression coefficients are significant.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

- Fit multiple linear regression models to a given set of data when using two or more predictor variables and perform residual analysis to check the validity of the models under consideration.
- Fit multiple linear regression models to a given set of data involving qualitative or categorical predictor variables.
- Determine the presence and possible elimination of multicollinearity.
- Use various criteria, such as the coefficient of multiple determination, adjusted coefficient of multiple determination, Mallows' C_p statistic, or PRESS statistics, to check the adequacy of the fitted model.
- Fit a logistic regression model when the response variable is a binary variable.
- Use Statistical packages MINITAB, R, and JMP to perform multiple regression analysis.

16.1 INTRODUCTION

In Chapter 15, we studied the simple linear regression model, which has one independent (predictor) variable. In practice, however, we deal more often with scenarios that have more than one independent variable. In other words, we use more than one predictor variable to explain the variation in the dependent variable Y. As a result we are usually in a better position to predict the dependent variable more accurately. For example, we may be interested in predicting the profit-sharing bonus for an employee in a given physical year. Clearly, many independent variables should be included, such as the company's revenues and profit for that physical year, employee's length of service at that company, his/her overall seniority, and productivity. As another illustration, suppose that we want to predict a person's total cholesterol level. Then, we may have to take into consideration independent variables such as weight, diet, age, health condition, and family history. Clearly, in both examples, if we have knowledge about all the independent variables, then we will certainly increase our ability to more accurately predict the profit-sharing bonus or the total cholesterol level.

In this chapter, we discuss regression models involving two or more independent variables, including models containing power terms such as $X_i, X_i^2, X_i^3, \ldots$, interaction terms such as $X_iX_j, X_iX_j^2, \ldots$, and models that have some qualitative independent variables. Regression models containing quadratic terms such as $X_i^2, X_j^2, X_iX_j, \ldots$, are called *second-order multiple linear regression models.* Such a model is called linear because it is a linear function of the regression coefficients. Finally, to select better-fitted regression models, we discuss residual analysis, detection of outliers, and testing of hypotheses for individual or group of regression coefficients, as well as other criteria such as the multiple coefficient of determination, R^2, adjusted R^2, multicollinearity, Mallows' C_p statistic, and the Prediction Error Sum of Squares (PRESS) statistic.

16.2 MULTIPLE LINEAR REGRESSION MODELS

We now consider the scenario in which the dependent variable Y can be explained only by two, and sometimes more independent variables. To introduce this type of scenario, we consider the following example.

Table 16.2.1 Students' placement scores and their cumulative GPAs.

X_1	84	87	92	84	94	88	86	89	95	85	80	91	85	81	87	85	80	82	85	90
X_2	86	85	88	82	93	87	88	90	92	87	84	90	88	83	83	85	82	83	80	86
Y	3.17	3.13	3.5	3.1	3.7	3.2	3.2	3.53	3.75	3.37	3.2	3.56	3.3	3.05	3.33	3.26	3.0	2.9	3.1	3.39

Example 16.2.1 (Placement test scores) *At a large university, a group of the faculty strongly believes that the success of students is closely related to their scores on the placement tests in mathematics and English, where the success of a student is measured by his/her cumulative GPA. The regression model that might explain the faculty's belief is*

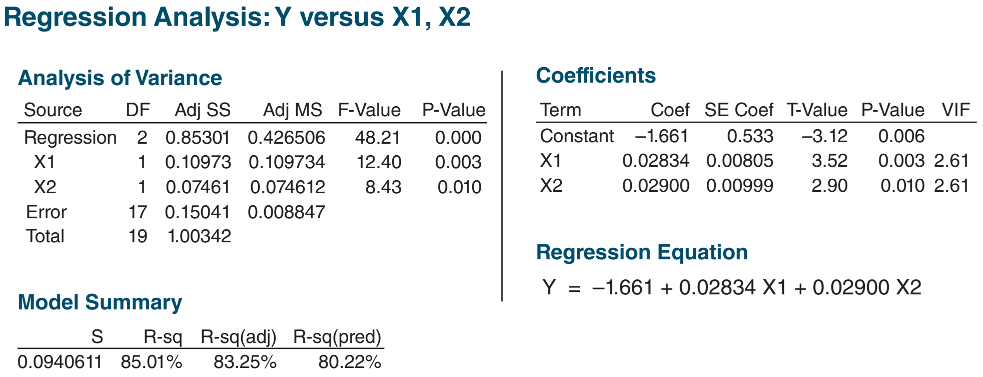
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{16.2.1}$$

where Y represents the cumulative GPA and X_1 and X_2 represents the scores on the placement tests in mathematics and English, respectively. In order to examine whether or not the scores on the placement tests determines the success of the students, the university authorities selected a random sample of 20 students and recorded their cumulative GPA and the placement scores as in Table 16.2.1.

MINITAB

To analyze these data using MINITAB, we proceed as follows:

1. Enter the values Y, X_1, and X_2 in columns C1, C2, and C3, respectively, and use Y, X_1, and X_2 as the headings of these columns.
2. From the Menu bar select **Stat** > **Regression** > **Regression** > **Fit Regression Model**. In the dialog box that appears, type Y in the box next to **Responses**, and type X_1 X_2 box next to **Continuous predictors**. Select any other desired options available in this dialog box and make necessary entries and click **OK**. The MINITAB output that appears in the Session window is shown below:



```
Regression Analysis: Y versus X1, X2

Analysis of Variance                                  Coefficients

Source      DF   Adj SS     Adj MS   F-Value  P-Value  Term       Coef     SE Coef  T-Value  P-Value  VIF
Regression   2  0.85301  0.426506     48.21    0.000   Constant  –1.661     0.533    –3.12    0.006
  X1         1  0.10973  0.109734     12.40    0.003   X1       0.02834  0.00805     3.52    0.003  2.61
  X2         1  0.07461  0.074612      8.43    0.010   X2       0.02900  0.00999     2.90    0.010  2.61
 Error      17  0.15041  0.008847
 Total      19  1.00342
                                                       Regression Equation

                                                       Y  =  –1.661 + 0.02834 X1 + 0.02900 X2
Model Summary

        S     R-sq  R-sq(adj)  R-sq(pred)
0.0940611   85.01%     83.25%      80.22%
```

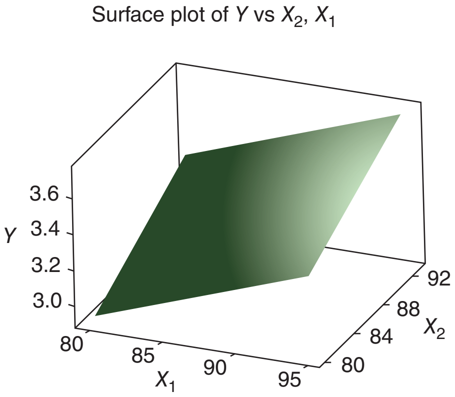


Figure 16.2.1 Response function surface plot (a plane) for the fitted regression model $\hat{Y} = -1.66 + 0.0283X_1 + 0.0290X_2$.

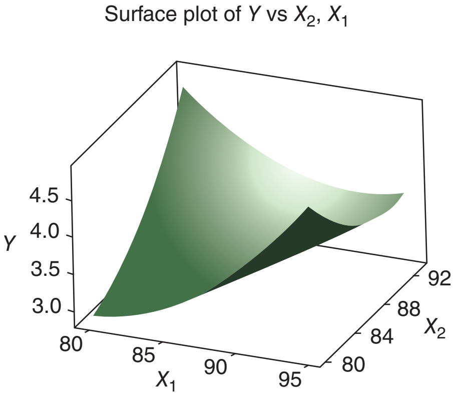
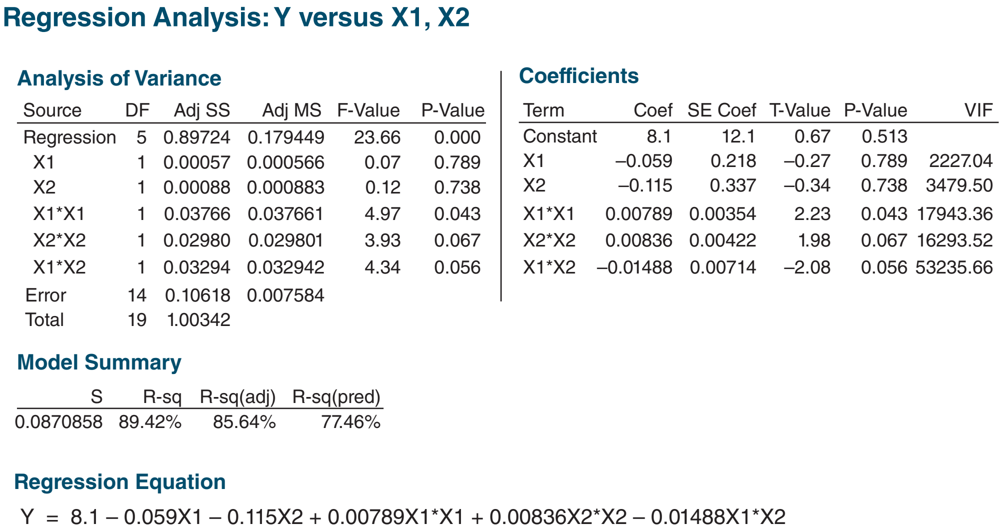


Figure 16.2.2 Response surface plot for the fitted regression model $\hat{Y} = 8.136 - 0.059X_1 - 0.115X_2 + 0.008X_1^2 + 0.008X_2^2 - 0.015X_1X_2$.

Note that the response surface in Figure 16.2.1 is a plane. This response function surface plot is obtained selecting the **Stat** > **Regression** > **Regression** > **Surface Plot**. When the dialog box appears, select variable *Y* as the Response. Then enter variable *X1* as the X axis and variable *X2* as the Y axis.

Now, if the model proposed contains interaction and quadratic terms, then the fitted response function surface is as shown in Figure 16.2.2. The model containing the interaction and quadratic terms is fitted using MINITAB by taking the same steps as given above, except that clicking on the **Model** button from the **Regression** window, allows for adding the necessary terms $(X1 * X1, X2 * X2, X1 * X2)$ when the new window appears. Then the following MINITAB output appears in the Session window.



```
Regression Analysis: Y versus X1, X2

Analysis of Variance                                        | Coefficients
Source      DF  Adj SS    Adj MS    F-Value  P-Value        | Term      Coef      SE Coef  T-Value  P-Value   VIF
Regression   5  0.89724   0.179449   23.66    0.000         | Constant   8.1      12.1       0.67    0.513
  X1         1  0.00057   0.000566    0.07    0.789         | X1        −0.059     0.218    −0.27    0.789   2227.04
  X2         1  0.00088   0.000883    0.12    0.738         | X2        −0.115     0.337    −0.34    0.738   3479.50
  X1*X1      1  0.03766   0.037661    4.97    0.043         | X1*X1      0.00789   0.00354   2.23    0.043  17943.36
  X2*X2      1  0.02980   0.029801    3.93    0.067         | X2*X2      0.00836   0.00422   1.98    0.067  16293.52
  X1*X2      1  0.03294   0.032942    4.34    0.056         | X1*X2     −0.01488   0.00714  −2.08    0.056  53235.66
Error       14  0.10618   0.007584
Total       19  1.00342

Model Summary
        S     R-sq  R-sq(adj)  R-sq(pred)
0.0870858   89.42%     85.64%      77.46%

Regression Equation
Y  =  8.1 – 0.059X1 – 0.115X2 + 0.00789X1*X1 + 0.00836X2*X2 – 0.01488X1*X2
```

Solution: USING R

To perform the required regression analysis on Y, we can use the following R-code. To include higher-order terms $X1 * X1, X2 * X2, X1 * X2$ in the regression model, we should add $I(X1^2) + I(X2^2) + I(X1 * X2)$ to the model as shown in model2 below.

```
Y = c(3.17,3.13,3.5,3.1,3.7,3.2,3.2,3.53,3.75,3.37,3.2,3.56,3.3,
3.05,3.33,3.26,3,2.9,3.1,3.39)
X1 = c(84,87,92,84,94,88,86,89,95,85,80,91,85,81,87,85,80,82,85,90)
X2 = c(86,85,88,82,93,87,88,90,92,87,84,90,88,83,83,85,82,83,80,86)


#Fitting MLR model using predictors X1 and X2
model1 = lm(Y ~ X1 + X2)
model1
anova(model1)
summary(model1)


#Fitting MLR model by adding both quadratic and interaction terms.
model2 = lm(Y ~ X1 + X2 +I(X1^2) + I(X2^2) + I(X1*X2))
model2
anova(model2)
summary(model2)
```

The results would be similar to those obtained using MINITAB.

We note that the analysis of Example 16.2.1 is carried out under the assumption that the random errors ε_i are independent $N(0, \sigma^2)$ random variables. Using this assumption in the ensuing sections, we discuss multiple regression models in more detail, including residual analysis and certain diagnostic tests, and illustrating them with data from Example 16.2.1.

We begin by considering a general regression model with k predictor variables, which may be stated as follows. A response variable Y is affected by the k predictor variables $(X_1, X_2, \dots, X_k)$ linearly, so that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{16.2.2}$$

where $X_1, X_2, \dots, X_k$ are k independent variables, $\beta_0, \beta_1, \dots, \beta_k$ are the regression coefficients and ε is a random error incurred when observing Y at $(X_1, X_2, \dots, X_k)$. Further, we assume that $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$, so that

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{16.2.2a}$$

Model (16.2.2) is called *linear* because it is a linear function in the regression parameters.

Recall that in Chapter 15 the simple linear regression model (15.2.1) represents a straight-line. Here, the multiple linear regression model (16.2.2) represents a hyperplane in the $(k + 1)$-dimensional space of $(X_1, X_2, \dots, X_k, Y)$. The regression coefficient β_j, represents the expected rate of change in Y as X_j changes when the remaining variables $X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_k$ are kept fixed; that is, as X_j changes by one unit, the dependent variable Y is expected to change by β_j units. Since β_j represents the expected change in the dependent variable Y as X_j changes by one unit when all other independent variables remain unchanged, the regression coefficients β_j's are also referred to as *partial regression coefficients.*

To analyze model (16.2.2), we assume that n observations are made on Y, say $Y_1, \dots, Y_i, \dots, Y_n$, where

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad n > k + 1 \tag{16.2.3}$$

In model (16.2.3) we assume the following:

1. Y_i is the value of the response or the dependent variable in the ith trial $(i = 1, 2, \dots, n)$.
2. $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are the unknown parameters, where β_0 is the intercept of the hyperplane and $\beta_1, \beta_2, \dots, \beta_k$ are the partial regression coefficients.
3. X_{ij} $(j = 1, 2, \dots, k)$ is the value of the predictor variable X_j in the ith trial.
4. ε_i is a random variable with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and ε_i and ε_j (for all $i, j; i \neq j$), $i, j = 1, 2, \dots, n$, are uncorrelated.

Furthermore, note that Y_i is a value of an observable random variable, the independent variables X_j are assigned at desired values; and ε_i is an unobservable random variable. Also $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are unknown parameters, that we want to estimate. The selection of the values of X_{ij} is called the *design of the experiment.*

As noted in Chapter 15, even when there is one independent variable, we encounter situations that fall into the same category as the multiple linear regression model. For example, if we fit a second-order regression model for one independent variable, that is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

it can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

by letting $X_2 = X_1^2$. In other words, any polynomial linear regression model (a second-order or higher-order linear regression model is called a *polynomial linear regression model*) can always be treated as a multiple linear regression model.

16.3 ESTIMATION OF REGRESSION COEFFICIENTS

The least-squares method used in Chapter 15 to fit a simple linear regression model is now extended to fit the multiple linear regression model (16.2.3). The least-squares method proceeds by minimizing the sum of squared deviations of the observed Y_i $(i = 1, 2, \ldots, n)$ from true means $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$, that is, by minimizing

$$Q(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_k X_{ik})^2 \tag{16.3.1}$$

over choices of $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$. The minimization of Q in (16.3.1) is achieved by taking the partial derivatives of Q with respect to $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and equating them to zero. When equating to zero, we denote the solution of the resulting equations by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$. Hence, we have the $(k + 1)$ equations

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\displaystyle\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \cdots - \hat{\beta}_k X_{ik}) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\displaystyle\sum_{i=1}^{n}(Y_i X_{i1} - \hat{\beta}_0 X_{i1} - \hat{\beta}_1 X_{i1}^2 - \hat{\beta}_2 X_{i2} X_{i1} - \cdots - \hat{\beta}_k X_{ik} X_{i1}) = 0 \\ \dfrac{\partial Q}{\partial \beta_2} = -2\displaystyle\sum_{i=1}^{n}(Y_i X_{i2} - \hat{\beta}_0 X_{i2} - \hat{\beta}_1 X_{i1} X_{i2} - \hat{\beta}_2 X_{i2}^2 - \cdots - \hat{\beta}_k X_{ik} X_{i2}) = 0 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \dfrac{\partial Q}{\partial \beta_k} = -2\displaystyle\sum_{i=1}^{n}(Y_i X_{ik} - \hat{\beta}_0 X_{ik} - \hat{\beta}_1 X_{i1} X_{ik} - \hat{\beta}_2 X_{i2} X_{ik} - \cdots - \hat{\beta}_k X_{ik}^2) = 0 \end{cases} \tag{16.3.2}$$

We have denoted the solutions to the $(k+1)$ Equations (16.3.2) by $\hat{\beta}_i$'s, which are called the least-squares estimators of the regression coefficients β_i's. We can also put Equations (16.3.2) into standard form, referred to as *normal equations*, as follows: expressions in the $\hat{\beta}_i$'s $(i = 0, 1, 2, \dots, k)$ appear on the left-hand sides and expressions in Y_i's appear on the right-hand sides of the equations. We then obtain the *normal equations*:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} X_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ik} = \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} \hat{\beta}_0 X_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{i2}X_{i1} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ik}X_{i1} = \sum_{i=1}^{n} Y_i X_{i1} \\ \sum_{i=1}^{n} \hat{\beta}_0 X_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1}X_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} X_{i2}^2 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ik}X_{i2} = \sum_{i=1}^{n} Y_i X_{i2} \\ \quad\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ \sum_{i=1}^{n} \hat{\beta}_0 X_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1}X_{ik} + \hat{\beta}_2 \sum_{i=1}^{n} X_{i2}X_{ik} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ik}^2 = \sum_{i=1}^{n} Y_i X_{ik} \end{cases} \tag{16.3.3}$$

Now, solving the system of $(k + 1)$ Equations (16.3.3) for the $k + 1$ unknowns $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, we obtain the least-squares estimators for $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. We illustrate the process of obtaining least-squares estimators with the following example.

Example 16.3.1 (Placement test scores) *Refer to Example 16.2.1. Find the least-squares estimates for $\beta_0, \beta_1,$ and β_2.*

Solution: Using the data of Table 16.2.1, we have the following $(n = 20,\ k + 1 = 3)$:

$$\sum_{i=1}^{20} Y_i = 65.74, \quad \sum_{i=1}^{20} X_{i1} = 1730, \quad \sum_{i=1}^{20} X_{i2} = 1722, \quad \sum_{i=1}^{20} X_{i1}^2 = 150{,}002,$$
$$\sum_{i=1}^{20} X_{i2}^2 = 148{,}496, \quad \sum_{i=1}^{20} X_{i1}X_{i2} = 149{,}179, \quad \sum_{i=1}^{20} Y_i X_{i1} = 5703.18, \quad \sum_{i=1}^{20} Y_i X_{i2} = 5673.34$$

Using the first three equations of (16.3.3) obtains the least-squares normal equations for model (16.2.1):

$$\begin{aligned} 20\hat{\beta}_0 + 1730\hat{\beta}_1 + 1722\hat{\beta}_2 &= 65.74 \\ 1730\hat{\beta}_0 + 150{,}002\hat{\beta}_1 + 149{,}179\hat{\beta}_2 &= 5703.18 \\ 1722\hat{\beta}_0 + 149{,}179\hat{\beta}_1 + 148{,}496\hat{\beta}_2 &= 5673.34 \end{aligned}$$

Solving these equations for $\hat{\beta}_0, \hat{\beta}_1,$ and $\hat{\beta}_2$, we have $\hat{\beta}_0 = -1.6609, \hat{\beta}_1 = 0.02834,$ $\hat{\beta}_2 = 0.02899$. These estimates of regression coefficients clearly match the results obtained by using MINITAB in Example 16.2.1. Note that most of the calculations can be done by using one of the statistical packages discussed in this book.

16.3.1 Estimation of Regression Coefficients Using Matrix Notation

The estimation process for estimating regression coefficients in model (16.2.2) is much simpler to describe using matrix notation. For example, in terms of matrices, model (16.2.2) can be expressed as

$$\mathrm{Y} = X\beta + \varepsilon \tag{16.3.4}$$

where

$$\underset{[n\times 1]}{\mathrm{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{[n\times(k+1)]}{\mathrm{X}} = \begin{bmatrix} 1 & X_{11} & X_{12}\cdots X_{1k} \\ 1 & X_{21} & X_{22}\cdots X_{2k} \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 1 & X_{n1} & X_{n2}\cdots X_{nk} \end{bmatrix}, \quad \underset{[(k+1)\times 1]}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \underset{[n\times 1]}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We say that $\beta_{(k+1)\times 1}$ is a vector of regression coefficients, and its least-squares estimator $\hat{\beta}$ minimizes the residual sum of squares Q, where

$$Q = (Y - X\beta)'(Y - X\beta) \tag{16.3.5}$$

Thus the least-squares estimator, $\hat{\beta}$, which is a $(k+1)\times 1$ vector, satisfies the normal equations written in matrix notation as

$$\frac{\partial}{\partial\beta}(Y - X\beta)'(Y - X\beta) = 0 \tag{16.3.6}$$

Using the partial differentiation indicated in (16.3.6), we obtain the solution $\hat{\beta}$, a $[(k+1)\times 1]$ vector, which is such that

$$-2X'Y + 2(X'X)\hat{\beta} = 0 \tag{16.3.7}$$

Thus, the standard form of the least-squares normal equations for model (16.3.4) is given by

$$(X'X)\hat{\beta} = X'Y \tag{16.3.8}$$

Now, assuming that X is a full rank matrix, the solution vector $\hat{\beta}$ is given by

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{16.3.9}$$

that is, the least-squares estimator of β is $\hat{\beta}$ as given by (16.3.9).

We illustrate the estimation of regression coefficients using matrix notation with the following example.

Example 16.3.2 (Placement test scores) *Refer to Example 16.2.1. Find the least square estimates for the regression coefficients of model (16.2.1) of Example 16.2.1 using matrix notation.*

Solution: In Example 16.2.1 we used model (16.2.1) with $n = 20$, $k + 1 = 3$, *or* $k = 2$. The model used to fit the data in Example 16.2.1 can also be written in matrix notation as

$$Y = X\beta + \varepsilon$$

where

$$\underset{[20\times1]}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{20} \end{bmatrix}, \quad \underset{[20\times3]}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{20,1} & X_{20,2} \end{bmatrix}, \quad \underset{[3\times1]}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \underset{[20\times1]}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{20} \end{bmatrix}$$

Substituting the values of the dependent variables Y_i and the predictor variables X_{ij}, we have

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 84 & 87 & \cdots & 85 & 90 \\ 86 & 85 & \cdots & 80 & 86 \end{bmatrix} \begin{bmatrix} 1 & 84 & 86 \\ 1 & 87 & 85 \\ \cdots & \cdots & \cdots \\ 1 & 85 & 80 \\ 1 & 90 & 86 \end{bmatrix} = \begin{bmatrix} 20 & 1730 & 1722 \\ 1730 & 150{,}002 & 149{,}179 \\ 1722 & 149{,}179 & 148{,}496 \end{bmatrix}$$

so that

$$(X'X)^{-1} = \begin{bmatrix} 32.0788 & -0.018692 & -0.353217 \\ -0.018692 & 0.0073177 & -0.0071346 \\ -0.353217 & -0.0071346 & 0.0112701 \end{bmatrix}$$

Further,

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 84 & 87 & \cdots & 85 & 90 \\ 86 & 85 & \cdots & 80 & 86 \end{bmatrix} \begin{bmatrix} 3.17 \\ 3.13 \\ \vdots \\ 3.10 \\ 3.39 \end{bmatrix} = \begin{bmatrix} 65.74 \\ 5703.18 \\ 5673.34 \end{bmatrix}$$

Thus, from Equation (16.3.9), we have that the estimates of β_0, β_1, and β_2 are such that

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 32.0788 & -0.018692 & -0.353217 \\ -0.018692 & 0.0073177 & -0.0071346 \\ -0.353217 & -0.0071346 & 0.0112701 \end{bmatrix} \begin{bmatrix} 65.74 \\ 5703.18 \\ 5673.34 \end{bmatrix} = \begin{bmatrix} -1.66092 \\ 0.02834 \\ 0.02899 \end{bmatrix}$$

Hence, the fitted least square regression model is

$$\hat{Y} = -1.66092 + 0.02834X_1 + 0.02899X_2$$

16.3.2 Properties of the Least-Squares Estimators

In this section, we study some properties of the estimator $\hat{\beta}$ under the conditions made for the regression model (16.2.3). Recall that model (16.2.3) assumed that ε_i is a random variable with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and that the ε_i's are uncorrelated, that is, for all $i \neq j$, $Cov(\varepsilon_i, \varepsilon_j) = 0$. We remark here that if the ε_i are independent, then the ε_i are uncorrelated, but the converse is not true; that is, if the ε_i are uncorrelated, it does not follow that they are independent. However, if we assume that ε_i are uncorrelated and have the $N(0, \sigma^2)$ distribution, then they are independent. Under the assumptions above (without normality) we state that

$$E(Y) = X\beta, \quad Var(Y) = Var(\varepsilon) = \sigma^2 I_n$$

where I_n is the identity matrix.

Now we show that:

1. $\hat{\beta}$ is an unbiased estimator of β, that is,

$$E(\hat{\beta}) = E((X'X)^{-1}X'Y) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta \tag{16.3.10}$$

2. The variance–covariance matrix of $\hat{\beta}$ is given by

$$\boxed{Var(\hat{\beta}) = (X'X)^{-1}\sigma^2 \qquad (16.3.11)}$$

To see this result we use the well-known theorem that says, if B is a matrix of constants and W is a random variable vector with variance–covariance matrix V_w then the random variable $U = BW$ has variance–covariance matrix given by

$$V(U) = BV_W B'$$

Now $\hat{\beta} = BY$, where $B = (X'X)^{-1}X'$ and $Var(Y) = \sigma^2 I_n$. Hence, for I_n denoting the $n \times n$ identity matrix, we have

$$\begin{aligned} Var(\hat{\beta}) &= (X'X)^{-1}X'(\sigma^2 I_n)((X'X)^{-1}X')' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{16.3.11a}$$

We could denote the matrix $(X'X)^{-1}$ in Equation (16.3.11) by A, say; then it would be easily seen that for each $i = 0, 1, 2, \ldots, k$,

$$Var(\hat{\beta}_i) = a_{ii}\sigma^2 \tag{16.3.12}$$

and

$$Cov(\hat{\beta}_i, \hat{\beta}_j) = a_{ij}\sigma^2 \quad \text{forall} \quad i \neq j \tag{16.3.13}$$

where a_{ij} is the $(i+1, j+1)$th element of the matrix A. As we noted in Chapter 15, to test hypotheses about the regression parameters and to find confidence intervals, we must first assume normality so that the ε_i are independently and identically distributed as $N(0, \sigma^2)$. Under normality, it can be easily shown that the estimator $\hat{\beta}$ is also a *maximum likelihood estimator* (MLE) of β. Finally, using an important theorem of statistics called the *Gauss–Markov Theorem*, we note that the least square estimator $\hat{\beta}$ is a linear unbiased estimator that attains minimum variance in the class of all linear unbiased estimators.

16.3.3 The Analysis of Variance Table

In the multiple linear regression case it can be shown, as in case of simple linear regression, that

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \tag{16.3.14}$$

That is,

Total sum of squares = Error sum of squares + Regression sum of squares

or

$$SS_{Total} = SSE + SSR \tag{16.3.14a}$$

from which we have

$$SSR = SS_{Total} - SSE \tag{16.3.14b}$$

In matrix notation, the sum of squares at the left side of Equation (16.3.14) can be rewritten as

$$SS_{Total} = Y'Y - \frac{1}{n}Y'JY = Y'\left(I - \frac{1}{n}J\right)Y \tag{16.3.15}$$

where I is an $(n \times n)$ identity matrix and J is an $(n \times n)$ matrix of 1's. Also the error sum of squares can be rewritten as

$$SSE = Y'Y - \hat{\beta}'X'Y \tag{16.3.16}$$

Now, substituting the value of $\hat{\beta}$ from (16.3.9), we obtain from (16.3.16) that

$$\begin{aligned} SSE &= Y'Y - ((X'X)^{-1}X'Y)'X'Y \\ &= Y'Y - Y'X(X'X)^{-1}X'Y \\ &= Y'(I - X(X'X)^{-1}X')Y \end{aligned} \tag{16.3.16a}$$

Finally, using (16.3.15) and (16.3.16), we can show that the regression sum of squares is expressible as

$$SSR = \hat{\beta}'X'Y - \frac{1}{n}Y'JY \tag{16.3.17}$$

Table 16.3.1 ANOVA table for the fitted regression model (16.2.2).

Source of variation	Sum of squares	Degree of freedom	Mean squares	F-ratio
Regression	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	k	$MSR = \dfrac{SSR}{k}$	$\dfrac{MSR}{MSE}$
Residual	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$n - (k+1)$	$MSE = \dfrac{SSE}{n-(k+1)}$	
Total	$SS_{Total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$	$n - 1$		

Table 16.3.2 ANOVA table in matrix notation for the fitted regression model (16.2.2).

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F-ratio
Regression	$SSR = Y'\left(X(X'X)^{-1}X' - \frac{1}{n}J\right)Y$	k	$MSR = \dfrac{SSR}{k}$	$\dfrac{MSR}{MSE}$
Residual	$SSE = Y'(1 - X(X'X)^{-1}X')Y$	$n - (k+1)$	$MSE = \dfrac{SSE}{n-(k+1)}$	
Total	$SST = Y'\left(1 - \frac{1}{n}J\right)Y$	$n - 1$		

An alternative form for *SSR* may be found using (16.3.9) and (16.3.17), that is,

$$SSR = Y'\left(X(X'X)^{-1}X' - \frac{1}{n}J\right)Y \tag{16.3.17a}$$

The foregoing results may be summarized in ANOVA tables as shown in Tables 16.3.1 and 16.3.2 given below.

Suppose now that Y_is are independent and normally distributed. Then, we use the test statistic $F = MSR/MSE$ to test the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_1 : \quad \text{not all} \quad \beta_i \quad (i = 1, 2, \ldots, k) \text{ are zero}$$

We note that under H_0 and normality of the ε_i it can be shown that *MSR* and *SSE* are independent random variables, with $MSR \sim \sigma^2\chi^2_k/k$ and $MSE \sim \sigma^2\chi^2_{n-k-1}/(n-k-1)$. It follows that under H_0, $F = MSR/MSE \sim F_{k,n-k-1}$. Hence, if the observed value of the test statistic $F = MSR/MSE > F_{k,n-(k+1);\alpha}$, then, the null hypothesis is rejected at the α level of significance. Otherwise, we do not reject the null hypothesis. Further note that

$$\hat{\sigma}^2 = S^2 = MSE = SSE/[n - (k+1)] \tag{16.3.18}$$

is an unbiased estimator of the error variance σ^2.

Now Equations (16.3.14) and (16.3.14a) show that the total variation, that is, the sum of squares SS_{Total}, is divided into two parts: the sum of squares due to regression *SSR* and the error sum of squares (often called the residual sum of squares) *SSE*. For a better fit of a model, we want the sum of squares due to regression to be as large as possible. To evaluate the fit of the model, a measure that is very frequently used is the *coefficient of determination*, defined (see Chapter 15) as

$$R^2 = \frac{SSR}{SS_{Total}} = 1 - \frac{SSE}{SS_{Total}} \tag{16.3.19}$$

We remark that R^2 is such that $0 \leq R^2 \leq 1$. The value of R^2 continues to increase as more and more predictor variables are included in the model, so it attains its maximum when all the k independent variables are in the model. In other words, R^2 does not attain an optimal value for any particular subset of predictor variables. Hence it does not help us to identify which subset of predictor variables will produce a better model. An alternative measure that is preferred over R^2 is the *adjusted coefficient of determination*, usually denoted by R_a^2, defined as

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)}\left(\frac{SSE}{SS_{Total}}\right) \tag{16.3.20}$$

Note that in R_a^2 an adjustment has been made for the number of predictor variables included in the model. As can easily be seen from (16.3.14b), R_a^2 will increase only if SSE decreases. This usually occurs when we include predictor variables that are useful in predicting the dependent variable. If a predictor variable does not contribute to predicting the dependent variable in the model, then, unlike R^2, the value of R_a^2 decreases. This is because if a predictor variable is not very useful, the decrease in *SSE* is not enough to offset the decrease in the denominator degrees of freedom, $n - (k + 1)$.

16.3.4 More Inferences about Regression Coefficients

In this section, we discuss the testing of various hypotheses and confidence intervals for regression parameters. Throughout this section, we assume that the error terms in model (16.2.3) are independently and identically distributed as $N(0, \sigma^2)$.

Test and Confidence Interval for an Individual Regression Parameter β_i, $i = 0, 1, 2, \ldots, k$

Using Equations (16.3.10) and (16.3.12) and assuming that the random errors are normally distributed with mean of 0 and variance σ^2, it can easily be shown that

$$\hat{\beta}_i \sim N(\beta_i, a_{ii}\sigma^2)$$

so that

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{a_{ii}}} \sim N(0, 1)$$

It can also be shown that $\hat{\beta}_i$s are independent of SSE under the assumptions of this section, so that

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{a_{ii}}} \sim t_{n-(k+1)}, \quad i = 1, 2, \ldots, k$$

Here, a_{ii} is the $(i + 1)$st diagonal element of $(X'X)^{-1}, i = 0, 1, \ldots, k$, and $\hat{\sigma} = \sqrt{MSE}$.

Hence, as can easily be seen, the confidence interval for β_i with confidence coefficient $(1 - \alpha)$ is given by

$$\left(\hat{\beta}_i \pm t_{n-(k+1);\frac{\alpha}{2}}\hat{\sigma}\sqrt{a_{ii}}\right) \tag{16.3.21}$$

A test involving β_i may be carried out as follows. Consider:

1. $H_0: \ \beta_i = 0$ versus $H_1: \ \beta_i \neq 0$.
2. Suppose the probability of type I error is set at α.
3. Test statistic $t = \dfrac{\hat{\beta}_i - 0}{\hat{\sigma}\sqrt{a_{ii}}} \sim t_{n-(k+1)}$, under H_0.
4. Reject the null hypothesis H_0 if $|t| \geq t_{n-(k+1);\alpha/2}$. Otherwise, do not reject the null hypothesis H_0.

Note that this test can also be carried out using the F test by using $F = t^2$. Under H_0, F is distributed as $F_{1,n-(k-1)}$. Thus we would reject H_0 if

$$F = \frac{\hat{\beta}_i^2}{MSE \times (a_{ii})} > F_{1,n-k-1;\alpha}$$

Tests and Confidence Intervals for Subsets of $r(r \leq k)$ Regression Coefficients

Simultaneous confidence intervals for r parameters using the Bonferroni method are given by

$$\hat{\beta}_i \pm t_{n-(k+1);\alpha/2r}\hat{\sigma}\sqrt{a_{ii}} \tag{16.3.22}$$

Note here the use of the $\alpha/2r$ point of the $t_{n-(k+1)}$ distribution. Tests for subsets of r regression parameters are carried out as follows:

Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{16.3.23}$$

where the labeling is such that the effect of $r \quad (2 \leq r \leq k)$ predictor variables $X_{k-(r-1)}, \ldots, X_{k-1}, X_k$ as measured by the r regression coefficients $\beta_{k-(r-1)}, \ldots, \beta_{k-1}, \beta_k$, is of interest. Here, as usual, we assume that the ε_i's are independent and identically distributed as $N(0, \sigma^2)$. Suppose that we wish to test the null hypothesis that

$$H_0 : \beta_{k-(r-1)} = \beta_{k-(r-2)} = \cdots = \beta_k = 0 \quad \text{versus} \quad H_1 : \text{ Not all } \ \beta_{k-(r-1)}, \ldots, \beta_k \text{ are zero} \tag{16.3.24}$$

Note that under the null hypothesis, model (16.3.23) reduces to

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-r} X_{i,k-r} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{16.3.25}$$

We call model (16.3.23) the *full model* and model (16.3.25) the *reduced model.* Further, we denote the *sum of squares due to regression* for the full and the reduced model by SSR_1 and SSR_2, respectively, and the *error sum of squares* by SSE_1 and SSE_2, respectively. Then the test statistic for testing the hypothesis (16.3.24) is

$$F = \frac{(SSR_1 - SSR_2)/r}{SSE_1/[n-(k+1)]} \tag{16.3.26}$$

Now an alternative statistic that can be used for testing the hypothesis (16.3.24) is

$$F^* = \frac{(SSE_2 - SSE_1)/r}{SSE_1/[n-(k+1)]} \tag{16.3.27}$$

where both statistics F and F^* are distributed as $F_{r,n-(k+1)}$. This is not surprising, since the statistics in (16.3.26) and (16.3.27) are algebraically equivalent because

$$SS_{Total} = SSR_1 + SSE_1 = SSR_2 + SSE_2 \tag{16.3.28}$$

We reject the null hypothesis H_0 at significance level α if $F \geq F_{r,n-(k+1);\alpha}$. Otherwise, we do not reject the null hypothesis H_0.

Example 16.3.3 (Placement test scores) *Refer to Example 16.2.1, using the data in Table 16.2.1.*

(a) Find 95% confidence intervals for β_1 and β_2.
(b) Test the hypothesis $H_0 : \beta_2 = 0 \quad versus \quad H_1 : \beta_2 \neq 0$. Use $\alpha = 0.05$.

Solution: (a) In this example we have $n = 20, k = 2,$ and $r = 1$. From the MINITAB printout in Example 16.2.1, we have

$$\hat{\sigma}^2 = MSE = 0.00885$$

and from Example 16.3.2, using data in Table 16.2.1, we have

$$\hat{\beta}_1 = 0.02834, \quad \hat{\beta}_2 = 0.02899$$

Since $n - (k + 1) = 20 - 3 = 17$, we use the critical value $t_{17;0.025} = 2.110$. Further, the X_1 and X_2 columns lead to the matrix $X'X$ whose inverse is

$$(X'X)^{-1} = \begin{bmatrix} 32.0788 & -0.018692 & -0.353217 \\ -0.018692 & 0.0073177 & -0.0071346 \\ -0.353217 & -0.0071346 & 0.0112701 \end{bmatrix}$$

From this inverse matrix we obtain

$$a_{11} = 0.0073177, \quad a_{22} = 0.0112701$$

Now using the result in Equation (16.3.21), we find the 95% confidence interval for

$$\begin{aligned}\beta_1 : (\hat{\beta}_1 \pm t_{17;0.025}\hat{\sigma}\sqrt{a_{11}}) &= (0.02834 \pm 2.110 \times \sqrt{0.00885} \times \sqrt{0.0073177}) \\ &= (0.02834 \pm 0.01698) = (0.01136, 0.04532) \\ \beta_2 : (\hat{\beta}_2 \pm t_{17;0.025}\hat{\sigma}\sqrt{a_{22}}) &= (0.02899 \pm 2.110 \times \sqrt{0.00885} \times \sqrt{0.0112701}) \\ &= (0.02899 \pm 0.02107) = (0.00792, 0.05006)\end{aligned}$$

(b) For a test of hypothesis involving β_2 we proceed as follows:

1. $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$.
2. P(Type I error) $= \alpha = 0.05$.
3. Observed value of the test statistic $T = (\hat{\beta}_2 - \beta_2)/(\hat{\sigma}\sqrt{a_{22}})$ under H_0 is

$$t = \frac{0.02899 - 0}{0.09407 \times 0.10616} = 2.903$$

which is greater than the critical value $t_{17;0.025} = 2.110$.

Hence, we reject the null hypothesis H_0. From the MINITAB printout in Example 16.2.1, the p-value for testing the hypothesis $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$, is 0.010, which leads to the same conclusion that we reject the null hypothesis at significance level $\alpha = 0.05$.

Tests for Subsets of r ($r \leq k$) Regression Parameters Using Matrix Notation

Here, we rewrite model (16.3.4) as

$$Y = X_1\eta_1 + X_2\eta_2 + \varepsilon \tag{16.3.29}$$

where

$$\mathrm{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \; X_{1(n\times(k-r+1))} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,k-r} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,k-r} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,k-r} \end{bmatrix}, \; \eta_{1((k-r+1)\times 1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-r} \end{bmatrix},$$

$$X_{2(n\times r)} = \begin{bmatrix} X_{1,k-(r-1)} & X_{1,k-(r-2)} & \cdots & X_{1k} \\ X_{2,k-(r-1)} & X_{2,k-(r-2)} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n,k-(r-1)} & X_{n,k-(r-2)} & \cdots & X_{nk} \end{bmatrix}, \; \eta_{2(r\times 1)} = \begin{bmatrix} \beta_{k-(r-1)} \\ \beta_{k-(r-2)} \\ \vdots \\ \beta_k \end{bmatrix}, \; \varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Expressing X and β in model (16.3.4) in terms of X_1, X_2 and η_1, η_2, respectively, we have

$$E(Y) = X\beta$$

where

$$X = [X_1 \vdots X_2\,] \quad \text{and} \quad \beta = \begin{bmatrix} \eta_1 \\ \cdots \\ \eta_2 \end{bmatrix}$$

Now we test the hypothesis

$$H_0 : \eta_2 = 0 \quad \text{versus} \quad H_1 : \eta_2 \neq 0 \tag{16.3.30}$$

From (16.3.17) it follows that

$$\begin{aligned} SSR_1 &= \hat{\beta}' X'Y - \tfrac{1}{n} Y'JY \\ &= ((X'X)^{-1} X'Y)' X'Y - \tfrac{1}{n} Y'JY \\ &= Y'X(X'X)^{-1} X'Y - \tfrac{1}{n} Y'JY \end{aligned}$$

Similarly it can be shown that

$$SSR_2 = Y'X_1(X_1'X_1)^{-1}X_1'Y - \frac{1}{n} Y'JY$$

From (16.3.16a) we have

$$SSE_1 = Y'Y - Y'X(X'X)^{-1}X'Y$$

From (16.3.26) the test statistic for testing the hypothesis (16.3.30) is thus given by

$$\begin{aligned} F &= \frac{(SSR_1 - SSR_2)/r}{SSE_1/[n-(k+1)]} \\ &= \frac{(Y'X(X'X)^{-1}X'Y - Y'X_1(X_1'X_1)^{-1}X_1'Y)/r}{SSE_1/[n-(k+1)]} \\ &= \frac{(Y'[X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1']Y)/r}{SSE_1/[n-(k+1)]} \end{aligned}$$

Under H_0, F is distributed as $F_{r,n-(k+1)}$, so we reject the null hypothesis H_0 if $F \geq F_{r,n-(k+1);\alpha}$. Otherwise, we do not reject the null hypothesis H_0. Another notation that is commonly used to denote $SSR_1 - SSR_2$ is

$$SSR_1 - SSR_2 = R(\beta_{k-(r-1)}, \beta_{k-(r-2)}, \ldots, \beta_k | \beta_0, \beta_1, \ldots, \beta_{k-r}) \tag{16.3.31}$$

Here R represents the increase in the regression sum of squares when the predictor variables $X_{k-(r-1)}, X_{k-(r-2)}, \ldots, X_k$ are added to a model involving $X_1, \ldots, X_{k-r}$ and the constant term. In general, the regression sum of squares due to $X_1, \ldots, X_k$ can be partitioned as

$$R(\beta_1, \ldots, \beta_k | \beta_0) = R(\beta_1|\beta_0) + R(\beta_2|\beta_0, \beta_1) + \cdots + R(\beta_k|\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1}) \tag{16.3.32}$$

where the terms on the right-hand side represent the increase in regression sum of squares when the predictor variable X_i $(i = 1, 2, \ldots, k)$ is added to the model involving $X_1, \ldots, X_{i-1}$ and the constant term.

For example, the regression sum of squares in (16.3.31) can be partitioned as

$$\begin{aligned} &R(\beta_{k-(r-1)}, \beta_{k-(r-2)}, \ldots, \beta_k | \beta_0, \beta_1, \ldots, \beta_{k-r}) \\ &= R(\beta_{k-(r-1)}|\beta_0, \beta_1, \ldots, \beta_{k-r}) + R(\beta_{k-(r-2)}|\beta_0, \beta_1, \ldots, \beta_{k-r}, \beta_{k-(r-1)}) \\ &\quad + \cdots + R(\beta_k|\beta_0, \beta_1, \ldots, \beta_{k-r}, \ldots, \beta_{k-1}). \end{aligned}$$

Confidence Interval for an Expected Response at X_0 with Confidence Coefficient $(1 - \alpha)$

Now we suppose that we are interested in $E(Y|X_0)$, where the $[1 \times (k+1)]$ vector $X_0 = (1, X_{10}, \ldots, X_{k0})$ contains assigned values of $(1, X_1, \ldots, X_k)$. From our earlier discussion it can easily be shown that the standard error of $\hat{Y}_0 = X_0\hat{\beta}$ at the given X-vector X_0, is estimated by

$$\hat{\sigma}_{\hat{y}_0} = \hat{\sigma}\sqrt{X_0(X'X)^{-1}X'_0} \tag{16.3.33}$$

Also $\hat{Y}_0$ is distributed as a normal variable, assuming that ε_i are normally distributed. This implies that

$$\frac{\hat{Y}_0 - E(Y|X_0)}{\hat{\sigma}\sqrt{X_0(X'X)^{-1}X'_0}} \tag{16.3.34}$$

is distributed as a t-distribution with $n-(k+1)$ degrees of freedom. Hence, from Equation (16.3.34) it follows immediately that a confidence interval for the *expected response* $E(Y|X_0)$ at X_0 with confidence coefficient $(1-\alpha)$ is given by

$$(\hat{Y}_0 \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{X_0(X'X)^{-1}X'_0}) \tag{16.3.35}$$

Prediction Interval for a Future Response at X_0 with Confidence Coefficient $(1 - \alpha)$

Proceeding in the same manner as in Chapter 15, it can be shown that a *prediction interval* for a future response at X_0 with confidence coefficient $(1-\alpha)$ is given by

$$(\hat{Y}_0 \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X'_0}) \tag{16.3.36}$$

Example 16.3.4 (Placement test scores) *Refer to Example 16.2.1. Using the data in Table 16.2.1, find a 95% confidence interval for* $E(Y|X_1 = 83, X_2 = 89)$ *and a 95% prediction interval for* $(Y|X_1 = 83, X_2 = 89)$.

Solution: In order to find a confidence interval for $E(Y|X_1 = 83, X_2 = 89)$ and prediction interval for $(Y|X_1 = 83, X_2 = 89)$, we first need to find the estimated response $\hat{Y}_0$ at $X_1 = 83, X_2 = 89$. From Example 16.3.1 we have

$$\hat{Y} = -1.66092 + 0.02834X_1 + 0.02899X_2$$

Now, substituting the values of $X_1 = 83$ and $X_2 = 89$, we have for $X_0 = (1,\ X_1,\ X_2) = (1,\ 83,\ 89)$ the estimated value of Y at X_0 is

$$\hat{Y}_0 = 3.27141$$

In Example 16.2.1, from the ANOVA table, we have $\hat{\sigma}^2 = MSE = 0.00885$. The degrees of freedom associated with the t-distribution in Equations (16.3.34) and (16.3.35) is $n-(k+1) = 20-(2+1) = 17$. Thus, from Equation (16.3.35), we obtain a 95% confidence interval for $E(Y|X_1 = 83, X_2 = 89)$ as

$$(3.27141 \pm 2.110\sqrt{0.00885} \times \sqrt{0.378849})$$

which, after some arithmetic, reduces to

$$(3.14923, 3.39359)$$

Similarly, using the result given by Equation (16.3.36), we obtain a 95% prediction interval for $(Y|X_1 = 83, X_2 = 89)$ as

$$(3.27141 \pm 2.110\sqrt{0.00885} \times \sqrt{1 + 0.378849})$$

Again, with some simplification, we obtain the prediction interval as

$$(3.03833, 3.50449)$$

Note: The confidence intervals above can be found using MINITAB. In order to do so, follow the steps described in Example 15.4.3.

PRACTICE PROBLEMS FOR SECTION 16.3

1. Set up a first-order multiple linear regression model in three predictor variables.
2. (a) Set up a second-order interaction multiple linear regression model in three predictor variables.
   (b) Set up a second-order complete multiple linear regression model in three predictor variables.
   (c) Set up the models in (a) and (b) using matrix notation.
3. In Problem 1, determine the normal equations and find the least square estimators for the regression coefficients $\beta_0, \beta_1, \beta_2$, and β_3.
4. A study was made of the relationship among skein strength (Y, in lb) of #225 cotton yarn, mean fiber length (X_1, in 0.01 in.) and fiber tensile strength (X_2, in 1000 psi). Twenty combinations of X_1 and X_2 values were used and Y observed at each of these combinations. The observations yield the following results (data from Duncan, 1958):

$$\sum_{i=1}^{20} Y_i = 1908, \quad \sum_{i=1}^{20} X_{i1} = 1540, \quad \sum_{i=1}^{20} X_{i2} = 1502, \quad \sum_{i=1}^{20} X_{i1}^2 = 119{,}797,$$
$$\sum_{i=1}^{20} X_{i2}^2 = 113{,}104, \quad \sum_{i=1}^{20} X_{i1}X_{i2} = 115,678.9, \quad \sum_{i=1}^{20} X_{i1}Y_i = 148,248.6,$$
$$\sum_{i=1}^{20} X_{i2}Y_i = 143{,}626, \quad \sum_{i=1}^{20} Y_i^2 = 184{,}766$$

Using the results above, show that the least-squares normal equations for the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad i = 1, 2, \ldots, 20$$

are given by

$$\begin{pmatrix} 20 & 1540 & 1502 \\ 1540 & 119{,}797 & 115,678.9 \\ 1502 & 115,678.9 & 113{,}104 \end{pmatrix} \times \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1908.0 \\ 148,248.6 \\ 143,626.0 \end{pmatrix}$$

5. (a) Use the least square normal equations in Problem 4 to estimate the regression coefficients $\beta_0, \beta_1, \beta_2$ for the regression plane $E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.
   (b) Use the result obtained in (a) to estimate the value of Y if a piece of yarn is such that $X_1 = 75$ and $X_2 = 70$.
6. In Problem 4 find:
   (a) An estimate of the variance σ^2. (*Hint*: $\hat{\sigma}^2 = MSE = SSE/(n-3)$.)
   (b) A 95% confidence interval for $E(Y|X_1 = 75, X_2 = 70)$, assuming that $\varepsilon_i' s$ are normally distributed with mean zero and variance σ^2.
7. The pull strength of a wire bond is an important characteristic. The table below gives information on pull strength Y, die height X_1, post height X_2, loop height X_3, wire length X_4, bond width on the die X_5, and bond width on the post X_6 (from Myers and Montgomery, 1995, used with permission).

Y	X_1	X_2	X_3	X_4	X_5	X_6
8.0	5.2	19.6	29.6	94.9	2.1	2.3
8.3	5.2	19.8	32.4	89.7	2.1	1.8
8.5	5.8	19.6	31.0	96.2	2.0	2.0
8.8	6.4	19.4	32.4	95.6	2.2	2.1
9.0	5.8	18.6	28.6	86.5	2.0	1.8
9.3	5.2	18.8	30.6	84.5	2.1	2.1
9.3	5.6	20.4	32.4	88.8	2.2	1.9
9.5	6.0	19.0	32.6	85.7	2.1	1.9
9.8	5.2	20.8	32.2	93.6	2.3	2.1
10.0	5.8	19.9	31.8	86.0	2.1	1.8
10.3	6.4	18.0	32.6	87.1	2.0	1.6
10.5	6.0	20.6	33.4	93.1	2.1	2.1
10.8	6.2	20.2	31.8	83.4	2.2	2.1
11.0	6.2	20.2	32.4	94.5	2.1	1.9
11.3	6.2	19.2	31.4	83.4	1.9	1.8
11.5	5.6	17.0	33.2	85.2	2.1	2.1
11.8	6.0	19.8	35.4	84.1	2.0	1.8
12.3	5.8	18.8	34.0	86.9	2.1	1.8
12.5	5.6	18.6	34.2	83.0	1.9	2.0

   (a) Fit a multiple linear regression model using X_2, X_3, X_4, and X_5, as the predictor variables.
   (b) Test for significance of regression using analysis of variance with $\alpha = 0.05$. What are your conclusions?
   (c) Use the model from (a) to find a 95% prediction interval for the pull strength when $X_2 = 20$, $X_3 = 30$, $X_4 = 90$, and $X_5 = 2.0$.
8. Fit a multiple linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon$, to the data in Problem 7.
9. Refer to Problem 8.
   (a) For the model in Problem 8, and assuming normality of the Y's, test the hypothesis $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ versus H_1 : Not all parameters $\beta_4, \beta_5, \beta_6$ are zero. Use $\alpha = 0.05$.
   (b) Find the p-value for the test in part (a).

10. Refer to Problem 7 part (c). Find a 95% confidence interval for $E(Y)$ and a 95% prediction interval for Y at $X_2 = 20$, $X_3 = 30$, $X_4 = 90$, and $X_5 = 2.0$, assuming normality of the Y's.
11. Runs were made at various conditions of saturation X_1 and transisomers X_2. The response SCI, denoted by Y, is given below for the corresponding conditions of X_1 and X_2.
    (a) Fit the regression model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ to the data below.
    (b) Test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ versus H_1 : Not both β_1, β_2 are zero. Use $\alpha = 0.05$.

X_1	0.04	0.04	0.04	0.04	0.20	0.20	0.20	0.20	0.38	0.38	0.38	0.38
X_2	100	110	120	130	100	110	120	130	110	110	120	130
Y	67.1	64.0	44.3	45.1	69.8	58.5	46.3	44.1	74.5	60.7	49.1	47.7

16.4 MULTIPLE LINEAR REGRESSION MODEL USING QUANTITATIVE AND QUALITATIVE PREDICTOR VARIABLES

So far we have only considered the case of quantitative predictor variables. In practice, however, it is quite common to have quantitative and qualitative predictor variables. In this section, we discuss cases where we have one or more qualitative predictor variables along with some quantitative variables. For instance, the board of directors of a big corporation wishes to develop a model to determine the salaries of their upper-level management. To achieve their goal, they consider similar companies and collect data on certain variables, which include the dependent variable Y (salaries) and predictor variables X_1 (quarterly revenues), X_2 (quarterly profit), X_3 (number of employees directly or indirectly under that manager), X_4 (number of years with the company), and some qualitative variables such as gender of the manager and highest degree earned (MS, MBA, or PhD).

16.4.1 Single Qualitative Variable with Two Categories

Suppose that we are interested in fitting a regression model in which, besides having a quantitative variable, we have one qualitative variable with two categories C_1 and C_2. In order to use one or more qualitative predictor variables in a regression model, we need to introduce variables called *dummy variables* that do not assume values on a continuous scale. Rather, such variables are assigned the value of 0 and 1. In other words, the values 0 and 1 merely represent the absence and presence of a particular category, respectively. We first illustrate the use of a dummy variable in a model that contains only one quantitative predictor variable and one qualitative variable.

For this situation the regression model may be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{16.4.1}$$

where X_{i2} is the dummy variable defined as follows:

$$X_{i2} = \begin{cases} 0, & \text{if category } C_1 \text{ is present (and category } C_2 \text{ is absent)} \\ 1, & \text{if category } C_2 \text{ is present (and category } C_1 \text{ is absent)} \end{cases}$$

We then have that the matrix X defined in Section 16.3.1 for model (16.4.1) is given by

$$X_{n\times 3} = \begin{bmatrix} 1 & X_{11} & 0 \\ 1 & X_{21} & 0 \\ \vdots & \vdots & \vdots \\ 1 & X_{n_1 1} & 0 \\ \vdots & \vdots & \vdots \\ 1 & \ldots & 1 \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & 1 \end{bmatrix}$$

Here n_1 observations are taken with C_1 present and C_2 absent and $n_2 = n - n_1$ observations are taken with C_2 present and C_1 absent, and X_{i1} is the ith value of the predictor variable used to generate $Y_i, i = 1, 2, \ldots, n \quad (n = n_1 + n_2)$. Hence, model (16.4.1) when the qualitative variable is of category C_1 becomes

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{16.4.2}$$

Now if the qualitative variable is of category C_2, the model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 + \varepsilon_i$$

which we rewrite as

$$Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + \varepsilon_i \tag{16.4.3}$$

Note that the only difference between models (16.4.2) and (16.4.3) is in the intercept term; that is, the change in the category of the qualitative variable results in a change in the dependent variable Y by a constant β_2. Graphically, the two models are represented by the two parallel lines in Figure 16.4.1.

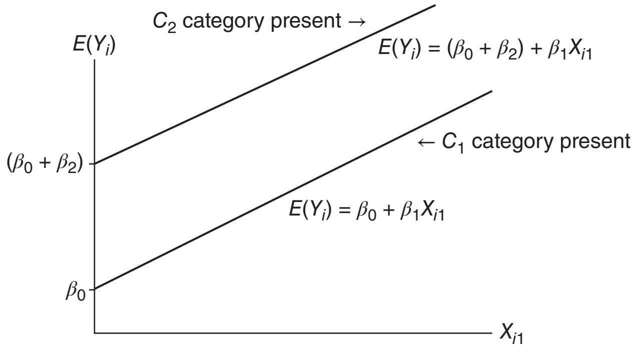


Figure 16.4.1 Graphical representation of models (16.4.2) and (16.4.3).

16.4.2 Single Qualitative Variable with Three or More Categories

We now consider a regression model with one quantitative predictor variable and one qualitative variable with three categories, say C_1, C_2, and C_3. In this case, the regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \tag{16.4.4}$$

where the predictor variables X_{i2}, X_{i3} are the *dummy variables* and are assigned the values 0 and 1 as follows:

		Assigned values	
		X_{i2}	X_{i3}
	C_1	1	0
Category present	C_2	0	1
	C_3	0	0

Hence, the models (16.4.4) for the presence of categories C_1, C_2, and C_3 (only one at a time) are given by

Only C_1 present

$$Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + \varepsilon_i \tag{16.4.5}$$

Only C_2 present

$$Y_i = (\beta_0 + \beta_3) + \beta_1 X_{i1} + \varepsilon_i \tag{16.4.6}$$

Only C_3 present

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{16.4.7}$$

respectively. The models (16.4.5)–(16.4.7) will result in the X matrix shown below for the case of n_j observations on Y taken with (only) category C_j present, $j = 1, 2, 3$, with $n_1 + n_2 + n_3 = n$:

$$X_{n\times 4} = \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n_1 1} & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{(n_1+1)1} & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & 0 & 1 \\ 1 & X_{(n_1+n_2)1} & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{(n_1+n_2+1)1} & 0 & 0 \\ 1 & \vdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & 0 & 0 \end{bmatrix}$$

Graphically, the models (16.4.5)–(16.4.7) can be represented by three parallel lines with intercepts $\beta_0 + \beta_2, \beta_0 + \beta_3$, and β_0, respectively. Note that the assertion of parallel lines is true only if the models are additive or there is no interaction between the quantitative and qualitative predictor variables. For example, suppose that the true situation calls for an interaction term between X_{i1} and X_{i2} but no interaction term between X_{i1} and X_{i3}, so that model (16.4.4) is not adequate. We have then that the true model would be

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \varepsilon_i \tag{16.4.8}$$

Hence, the model for category C_1 is given by

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_{i1} + \varepsilon_i \tag{16.4.9}$$

which obviously has a different slope than from that of the models for categories C_2 and C_3.

Example 16.4.1 (Corporate management salaries) *Consider the data in Table 16.4.1, generated from a study of salaries (Y in hundreds of thousands of dollars) of upper-level corporate management. In this study, we have four quantitative predictor variables: X_1 (quarterly revenues in billions of dollars), X_2 (quarterly profit in hundred million dollars), X_3 (number of employees in thousands, directly or indirectly under that manager), X_4 (number of years with the company) and two different types of qualitative predictor variables, gender and highest degree. We then assign X_5 to denote gender and X_6, X_7 to be used in concert for highest degree, as follows:*

		X_{i5}	
Gender categories	C_1 (Male)	1	
	C_2 (Female)	0	
		X_{i6}	X_{i7}
	D_1 (MS)	1	0
Highest degree categories	D_2 (MBA)	0	1
	D_3 (PhD)	0	0

We remark that the two qualitative factors on hand have different numbers of categories—the rule for the number of independent variables needed in the model to explain each qualitative factor is

$$\text{Number of independent variables} = \text{Number of categories} - 1$$

This explains why we need one variable (X_5) to describe gender and two variables (X_6, X_7) to describe the highest degree. Using this, we then have the data as given in Table 16.4.1.

Using the data in Table 16.4.1, we want to fit the multiple linear regression model

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \varepsilon_i, \; i = 1, 2, \ldots, n$$

Solution: The model that we use to analyze the data in Table 16.4.1 is

$$Y = X\beta + \varepsilon \tag{16.4.10}$$

Table 16.4.1 Salary data for upper-level management.

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
4.54	8.33	3.85	4.91	19	1	0	1
4.08	6.06	4.07	3.16	25	1	0	1
5.10	7.97	3.89	3.99	18	1	0	1
4.49	5.27	4.45	4.91	18	1	0	1
5.71	5.08	3.28	4.88	20	1	0	1
5.38	8.08	4.29	4.76	23	1	0	1
4.06	8.55	4.86	4.72	23	1	0	1
5.59	7.90	3.11	3.68	23	1	0	1
5.73	8.20	4.31	4.75	22	1	0	1
4.33	6.76	4.14	5.00	23	1	0	1
4.49	7.44	3.05	3.06	18	1	0	1
4.76	9.64	3.55	4.30	24	1	0	1
5.23	5.62	4.06	4.45	20	1	0	0
4.82	5.25	4.89	3.61	19	0	0	0
5.82	5.24	4.38	3.15	22	0	0	0
5.65	8.36	3.97	4.47	20	0	0	0
4.61	9.64	4.67	4.71	20	0	1	0
5.11	6.44	3.79	3.73	16	0	1	0
4.44	6.92	3.67	3.49	24	0	1	0
4.17	9.90	3.36	4.12	25	0	1	0

where

$$Y_{20\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{20} \end{bmatrix}, \quad X_{20\times 8} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots X_{17} \\ 1 & X_{21} & X_{22} & \cdots X_{27} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{20,1} & X_{20,2} & \cdots X_{20,7} \end{bmatrix},$$

$$\beta_{8\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_7 \end{bmatrix}, \quad \varepsilon_{20\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{20} \end{bmatrix}$$

Using Equation (16.4.10) and recalling Equation (16.3.8) that

$$(X'X)\hat{\beta} = X'Y$$

we find for the data in Table 16.4.1, the least-square normal equations for model (16.4.10) given by

$$\begin{pmatrix} 20.00 & 146.65 & 79.64 & 83.85 & 422.00 & 13.00 & 4.00 & 12.00 \\ 146.65 & 1122.03 & 581.60 & 619.57 & 3119.09 & 94.90 & 32.90 & 89.28 \\ 79.64 & 581.60 & 322.67 & 335.76 & 1678.73 & 50.91 & 15.49 & 46.85 \\ 83.85 & 619.57 & 335.76 & 359.82 & 1767.48 & 56.57 & 16.05 & 52.12 \\ 422.00 & 3119.09 & 1678.73 & 1767.48 & 9036.00 & 276.00 & 85.00 & 256.0 \\ 13.00 & 94.90 & 50.91 & 56.57 & 276.00 & 13.00 & 0.00 & 12.00 \\ 4.00 & 32.90 & 15.49 & 16.05 & 85.00 & 0.00 & 4.00 & 0.00 \\ 12.00 & 89.28 & 46.85 & 52.12 & 256.00 & 12.00 & 0.00 & 12.00 \end{pmatrix} \times \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_7 \end{bmatrix}$$

$$= \begin{pmatrix} 98.11 \\ 715.54 \\ 389.97 \\ 411.57 \\ 2063.47 \\ 63.49 \\ 18.33 \\ 58.26 \end{pmatrix}$$

Solving these normal equations, we obtain the solution $\hat{\beta}$ given by

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_7 \end{bmatrix} = \begin{pmatrix} 20.00 & 146.65 & 79.64 & 83.85 & 422.00 & 13.00 & 4.00 & 12.00 \\ 146.65 & 1122.03 & 581.60 & 619.57 & 3119.10 & 94.90 & 32.90 & 89.28 \\ 79.64 & 581.60 & 322.67 & 335.76 & 1678.73 & 50.91 & 15.49 & 46.85 \\ 83.85 & 619.57 & 335.76 & 359.82 & 1767.48 & 56.57 & 16.05 & 52.12 \\ 422.0 & 3119.09 & 1678.73 & 1767.48 & 9036.00 & 276.00 & 85.0 & 256.0 \\ 13.00 & 94.90 & 50.91 & 56.57 & 276.00 & 13.00 & 0.00 & 12.00 \\ 4.00 & 32.90 & 15.49 & 16.05 & 85.00 & 0.00 & 4.00 & 0.00 \\ 12.00 & 89.28 & 46.85 & 52.12 & 256.00 & 12.00 & 0.00 & 12.00 \end{pmatrix}^{-1}$$

$$\times \begin{pmatrix} 98.11 \\ 715.54 \\ 389.97 \\ 411.57 \\ 2063.47 \\ 63.49 \\ 18.33 \\ 58.26 \end{pmatrix}$$

$$= \begin{pmatrix} 7.25385 & -0.03320 & -0.47342 & 0.27406 & -0.02714 & -0.59201 & -1.08798 & -0.32294 \end{pmatrix}'$$

Using the solution above, we see that the $\hat{\beta}_i$s are such that the fitted regression model is

$$\begin{aligned} \hat{Y} = {} & 7.25385 - 0.03320X_1 - 0.47342X_2 + 0.27406X_3 - 0.02714X_4 - 0.59201X_5 \\ & - 1.08798X_6 - 0.32294X_7 \end{aligned}$$

This regression model can be used to estimate and/or predict the salary of an upper-level manager for given values of the predictor variables. The reader should keep in mind that the values of the predictor variables for which we want to predict the salary of a manager should fall within the range of the predictor variables used to fit this model. Suppose that we want to estimate or predict an observation generated at the "place" that the first observation is generated. We proceed, using

$$\hat{Y}_1 = 7.25385 - 0.03320X_1 - 0.47342X_2 + 0.27406X_3 - 0.02714X_4 - 0.59201X_5 \\ - 1.08798X_6 - 0.32294X_7$$

along with the information in row one of Table 16.4.1, so that the estimate of Y_1 is given by

$$\hat{Y}_1 = 7.25385 - 0.03320(8.33) - 0.47342(3.85) + 0.27406(4.91) - 0.02714(19) \\ -0.59201(1) - 1.08798(0) - 0.32294(1) = 5.06965 \approx 5.07$$

From Table 16.4.1 the observed value of Y_1 is 4.54. Thus, the error (residual) corresponding to the first observation is $e_1 = Y_1 - \hat{Y}_1 = 4.54 - 5.07 = -0.53$.

Table 16.4.2 gives all 20 fitted values, their standard errors, and the corresponding residuals.

Table 16.4.2 Observations, fitted values, standard error and residuals for Example 16.4.1.

Observations	Y_i	$\hat{Y}_i$	$SE\ (\hat{Y}_i)$	Residual
1	4.540	5.070	0.263	−0.530
2	4.080	4.398	0.404	−0.318
3	5.100	4.838	0.303	0.262
4	4.490	4.914	0.348	−0.424
5	5.710	5.412	0.447	0.298
6	5.380	4.720	0.228	0.660
7	4.060	4.424	0.339	−0.364
8	5.590	4.989	0.293	0.601
9	5.730	4.731	0.219	0.999
10	4.330	4.901	0.288	−0.571
11	4.490	4.998	0.424	−0.508
12	4.760	4.865	0.299	−0.105
13	5.230	5.230	0.592	−0.000
14	4.820	5.238	0.387	−0.418
15	5.820	5.273	0.388	0.547
16	5.650	5.779	0.453	−0.129
17	4.610	4.383	0.407	0.227
18	5.110	4.746	0.420	0.364
19	4.440	4.504	0.378	−0.064
20	4.170	4.697	0.402	−0.527

We can proceed to find the confidence interval for $E(Y|X = X_0) = X_0\beta$, where X_0 is a $[1 \times (k + 1)]$ vector of given values of predictor variables:

$$X_0 = (1, X_{01}, \ldots, X_{0k}) \tag{16.4.11}$$

so that X_0 is a row vector of order $(1 \times (k + 1))$. Note that β is a column vector of unknown regression coefficients and is of order $((k + 1) \times 1)$. We have that $\hat{Y}_0 = X_0\hat{\beta}$ is unbiased for $E(Y|X = X_0) = X_0\beta$. In order to find the confidence interval for $E(Y|X_0)$, we first need to find $Var(\hat{Y}|X_0)$, which is given by

$$Var(\hat{Y}|X_0) = \sigma^2(X_0(X'X)^{-1}X_0') \tag{16.4.12}$$

Then, the confidence interval for $E(Y|X_0)$ is given by

$$(\hat{Y}(X_0) \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'}) \tag{16.4.13}$$

where $\hat{\sigma} = \sqrt{MSE}$ and $\hat{Y}(X_0) = X_0\beta'$. Note that the quantity $\hat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'}$ in Equation (16.4.13) is an estimate of the standard error of $(\hat{Y}|X_0)$. As mentioned earlier, the standard errors of $(\hat{Y}|X_0)$, as we let X_0 take on values of $(1, X_{01}, \ldots, X_{07})$, $i = 1, 2, \ldots, 20$, of Table 16.4.1, are given in Table 16.4.2. Thus, in Example 16.4.1, a 95% confidence interval for $E(Y|X_0)$, when $X_0 = (1, X_{01}, \ldots, X_{07}) = (1, 8.33, 3.85, 4.91, 19, 1, 0, 1)$, which is used to generate Y_1 (see Table 16.4.1), is given by

$$\begin{aligned} \hat{Y}(X_0) \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'} &= 5.070 \pm t_{12;0.025}(0.263) \\ &= 5.070 \pm 2.179(0.263) \\ &= (4.497, 5.643) \end{aligned}$$

Following the results obtained in Chapter 15, the prediction interval for a "new" observation at a given value of X_0 is given by

$$\hat{Y}(X_0) \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'} \tag{16.4.14}$$

In general, if we choose X_0 to be any row of the X-matrix used to generate Y_i (above X_0 was the first row of the X-matrix), then using the notation of the HAT matrix (defined below in (16.4.15), the confidence intervals in (16.4.13) and (16.4.14) can be written as

$$\hat{Y}(X_0) \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{h_{ii}} \tag{16.4.14a}$$

$$\hat{Y}(X_0) \pm t_{n-(k+1);\alpha/2}\hat{\sigma}\sqrt{1 + h_{ii}} \tag{16.4.14b}$$

respectively, where h_{ii} is the ith diagonal element of the hat matrix H, which is defined as

$$H = X(X'X)^{-1}X' \tag{16.4.15}$$

Using the HAT matrix, the predicted values $\hat{Y}$ in terms of the observed values Y can be written as

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY \tag{16.4.16}$$

We remark that the residuals can be written in terms of the HAT matrix as

$$e = Y - \hat{Y} = Y - HY = (I - H)Y \tag{16.4.17}$$

where e is vector of residuals. Note that the matrices H and $(I - H)$ are *idempotent* (matrix B is called *idempotent* if and only if $B^2 = B$). From Equation (16.4.17) and the fact that $I - H$ is idempotent, we can easily see that the variance–covariance matrix of the residuals e can be written as

$$Cov(e) = \sigma^2(I - H) \tag{16.4.18}$$

We then have that

$$Var(e_i) = \sigma^2(1 - h_{ii}) \tag{16.4.19}$$

where h_{ij} $(i, j = 1, 2, \ldots, n)$ is the *ij*th entry of the HAT matrix *H*.

Note: Using MINITAB, the regression model in Example 16.4.1 can be fitted without creating any dummy variables. That is, enter the data in columns C1–C7 and name them Y, X1, X2, . . . , X5, and X6. The entries in column C6 are M and F for gender and in columns C7 the highest degrees. From the Menu bar select **Stat** > **Regression** > **Regression** > **Fit Regression Model**. In the dialog box that appears, enter *Y* in the box for **Response:** X1, X2, X3, and X4 in the box below **Continuous predictors:**, and X5, X6, and X7 in the box below **Categorical Predictors:**. Select other desired options from Graphs, Options, Storage, Results, and make the necessary entries, and click **OK**. Then the MINITAB output will appear in the Session window. We further illustrate the use of MINITAB with the following example.

Example 16.4.2 (Home prices) *During the 2008–2009 recession, home prices declined significantly throughout the United States. Many locations, however, were affected more seriously than others. The Orlando area in Florida was one location hit especially hard. Table 16.4.3 provides the listed prices for 28 randomly selected homes and associated predictor variables. Using MINITAB and R, fit a first-order multiple regression model with all the terms to these data. We let*

Y : *Listed price rounded in thousands of dollars*

Table 16.4.3 Listed selling prices of 28 randomly selected homes.

Y	X_1	X_2	X_3	X_4	Y	X_1	X_2	X_3	X_4
160	19	3	2	2	190	22	4	2	2
136	17	3	2	2	209	24	4	2	2
189	22	4	2	2	200	20	4	2	2
116	12	2	1	1	150	18	3	2	2
136	19	4	2	1	118	12	2	1	2
255	24	5	3	3	135	17	3	2	2
209	18	4	2	3	230	22	4	3	3
135	15	2	2	1	155	18	3	2	2
265	28	4	3	3	118	12	2	1	2
115	12	2	1	1	159	17	3	2	2
114	18	2	2	2	199	21	3	2	2
219	21	4	3	2	200	23	4	2	2
299	26	5	3	3	260	24	5	3	3
200	21	4	2	2	210	21	4	2	2

X_1 : *Living area rounded in hundreds of square feet*
X_2 : *Number of bedrooms*
X_3 : *Number of bathrooms*
X_4 : *Garage size (number of cars)*

MINITAB

To fit a regression model using MINITAB, we proceed as follows:

1. Enter the data in columns C1–C5. (C1 is a column of the 28 observations, etc.)
2. Select **Stat** > **Regression** > **Regression** > **Fit Regression Model**. A dialog box "General Regression" appears. In this dialog box, enter Y in the box for **Response** and X1, X2, X3, and X4 in the box below **Continuous predictors:**. Select other desired options from Graphs, Options, Storage, Results, and make the necessary entries; then click **OK**. Then the MINITAB output appears in the Session window as shown here.

```
Model Summary

      S     R-sq  R-sq(adj)  R-sq(pred)
18.4092   89.23%     87.35%      82.15%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value    VIF
Constant   –26.0     17.3    –1.51    0.146
X1          4.06     1.96     2.08    0.049   5.51
X2         19.96     7.59     2.63    0.015   4.22
X3           6.6     11.6     0.57    0.577   3.93
X4         22.50     8.41     2.68    0.014   2.06


Regression Equation

Y  =  –26.0 + 4.06X1 + 19.96X2 + 6.6X3 + 22.50X4


Analysis of Variance

Source         DF   Adj SS    Adj MS  F-Value  P-Value
Regression      4  64554.3   16138.6    47.62    0.000
  X1            1   1463.2    1463.2     4.32    0.049
  X2            1   2345.8    2345.8     6.92    0.015
  X3            1    108.8     108.8     0.32    0.577
  X4            1   2425.2    2425.2     7.16    0.014
Error          23   7794.7     338.9
  Lack-of-Fit  15   7350.0     490.0     8.82    0.002
  Pure Error    8    444.7      55.6
  Total        27  72349.0


Fits and Diagnostics for Unusual Observations

Obs         Y      Fit    Resid   Std Resid
  5    136.00   166.70   –30.70       –2.10   R
```

This printout gives the fitted model, and various t-tests showing that X_1, X_2, and X_4 are influential predictor variables. The values of R^2 and R^2_{adj} are reasonably high, which indicates that the overall model is good. However, as mentioned in Chapter 15, before making any final conclusion, we must check the adequacy of the model. The MINITAB graphs in Figure 16.4.2 show no abnormalities; that is, the assumptions of independence, normality, and constant variance are fairly satisfactory.

The analysis of variance table shows the overall fit is quite good since the p-value is zero. It also shows that lack of fit is significant. This means the present model is good, but not adequate; that is, we need to include either another factor or some other terms such as interaction terms. Note that the data contain 13 observations that are not replicated and 15 replicated observations (replicated in groups of 3, 2, 2, 2, 2, 2, 2). These are taken at $(X_1, X_2, X_3, X_4) = (17, 3, 2, 2)$, $(22, 4, 2, 2)$, $(17, 3, 2, 2)$, $(12, 2, 1, 2)$, $(24, 5, 3, 3)$, $(18, 3, 2, 2)$, $(21, 4, 2, 2)$. The pure error degrees of freedom come from the replicated observations $((3 - 1) + (2 - 1) + (2 - 1) + (2 - 1) + (2 - 1) + (2 - 1) + (2 - 1) = 8)$; that is, each replicated observation provides information about the pure error with number of replications minus one degrees of freedom so that in this case we have 8 degrees of

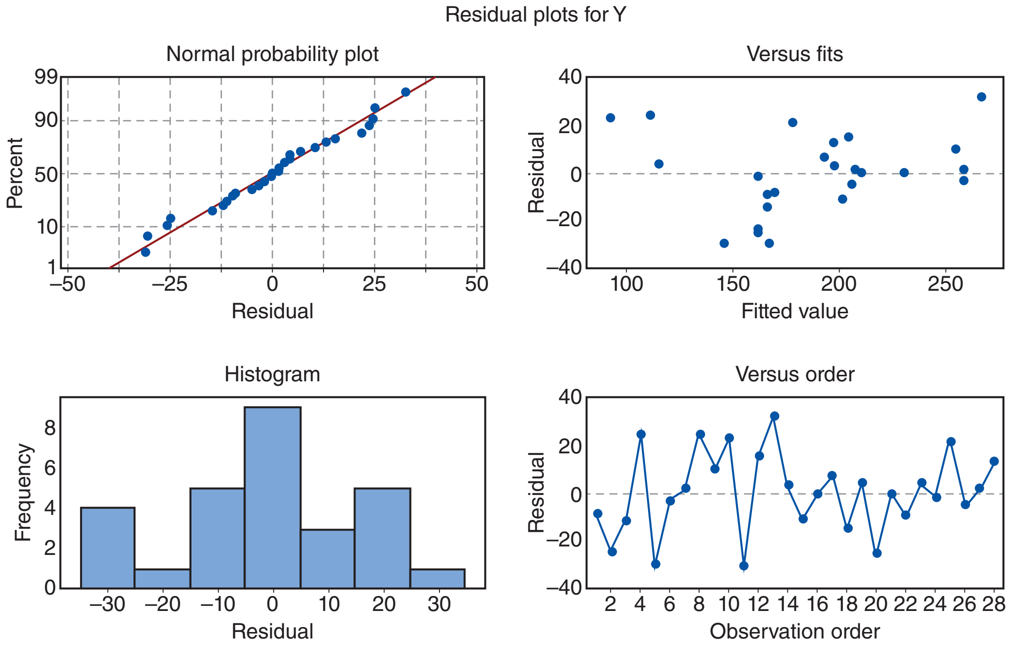


Figure 16.4.2 MINITAB printout of residual plots for data in Table 16.4.2.

freedom for the pure error. (See the analysis of variance table above.) Further note that the overall model is tested by using the regression mean sum of squares and the total residual mean sum of squares (the total residual sum of squares is the sum of squares due to *lack of fit* and *pure error*). The lack of fit is tested by using the lack-of-fit mean sum of square and the pure error mean sum of squares. In this example, the overall model and lack of fit are significant, since p-values are 0.000 and 0.002, respectively.

The MINITAB printout above provides the fit and the residual for the only extreme observation presence in the data based on the standardized residual calculation. For this observation, the listed value seems unusually lower compare to predicted value. The low listed price may be due to various reasons, such as the overall condition of the property may not be good, or the owner listed it low for a fast sale.

Solution: USING R To perform the required regression analysis on Y, we can use the following R-code.

```
Y = c(160,136,189,116,136,255,209,135,265,115,114,219,299,200,190,
209,200,150,118,135,230,155,118,159,199,200,260,210)
X1 = c(19,17,22,12,19,24,18,15,28,12,18,21,26,21,22,24,20,18,12,
17,22,18,12,17,21,23,24,21)
X2 = c(3,3,4,2,4,5,4,2,4,2,2,4,5,4,4,4,4,3,2,3,4,3,2,3,3,4,5,4)
X3 = c(2,2,2,1,2,3,2,2,3,1,2,3,3,2,2,2,2,2,1,2,3,2,1,2,2,2,3,2)
X4 = c(2,2,2,1,1,3,3,1,3,1,2,2,3,2,2,2,2,2,2,2,3,2,2,2,2,2,3,2)
```

```
#Fitting MLR model using predictors X1, X2, X3, and X4
model = lm(Y ∼ X1 + X2 + X3 + X4)
model
anova(model)
summary(model)
```

The results (not shown here) are similar to those obtained using MINITAB.

PRACTICE PROBLEMS FOR SECTION 16.4

1. A company has two water testing labs, Lab A and Lab B. Lab A is fully equipped with modern facilities, whereas Lab B does not have all those facilities. The following data give the number of water samples X_1 arriving per day in each of the two labs and the total of technician hours Y taken by the technicians to analyze all the samples arriving in that lab. The CEO of the company wants to fit a regression model to evaluate whether there is a benefit in upgrading Lab B to have the same facilities as Lab A. Fit an appropriate model to the data given below.

Y	78	86	80	92	60	78	72	80	82	98
X_1 (Water)	28	27	30	24	36	28	35	30	32	30
X_2 (Lab)	A	B	A	B	A	B	A	B	A	B

2. Refer to Problem 1 above. Suggest to the CEO how she can use this model to achieve the company's goal.
3. An educator is interested in studying the difference between public and private four-year institutions that award a chemical engineering degree. The educator selected eight public institutions and eight private institutions. The following data give the number of graduates (Y) hired during campus interviews, the number of students X_1 in the graduating class with a chemical engineering degree, and the type of institution X_2. Note that we have set $X_2 = 1$ for each public institution and $X_2 = 0$ for each private institution. Fit an appropriate linear regression model to these data.

Y	25	32	34	30	35	26	26	33	21	23	26	28	23	28	25	21
X_1	57	56	42	59	51	42	43	59	73	53	79	59	61	69	68	69
X_2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0

4. Refer to Problem 3 above. Develop the ANOVA table for the model you considered in Problem 3, and use an appropriate F-test to evaluate whether or not the model you fitted is appropriate. Use $\alpha = 0.05$.

5. Refer to Table 16.4.3 in Example 16.4.2. As an addendum to the data of Table 16.4.3, incorporate an additional qualitative predictor variable to be called "house siding" having four categories, namely Wood, Vinyl, Stucco, and Brick. Specifically, the first seven homes used wood siding, next seven used vinyl, the next seven used stucco, and the last seven homes used brick siding. Re-analyze the data in Table 16.4.3 taking into consideration the new information.
6. Refer to Problem 5 above. Determine a 95% confidence interval for $E(Y|X_0)$ and a prediction interval for Y when $X_0 = (25, 4, 3, 2, Brick)$.
7. Refer to Problem 5 above. Develop the ANOVA table for the model considered in Problem 5, and use an appropriate F-test to evaluate whether or not the model fitted is appropriate. Use $\alpha = 0.01$.
8. Refer to Problem 3 above. Determine a 95% confidence interval for $E(Y|X_0)$ and a prediction interval for Y when $X_0 = (50, 1)$ and $X_0 = (50, 0)$.

16.5 STANDARDIZED REGRESSION COEFFICIENTS

In a multiple linear regression model, it is difficult to compare the change in the response variable Y, if one predictor variable changes and the other predictor variables are held constant, because of the difference in the units of the predictor variables. For example, consider a fitted regression model

$$\hat{Y} = 40 + 500X_1 + 0.7X_2$$

where X_1 is measured in kilograms and X_2 is measured in grams. Obviously, the rate of change in the response variable is 500 units as X_1 changes one unit (i.e., 1 kg) while X_2 is held constant, whereas the rate of change in the response variable is only 0.7 units as X_2 changes 1 unit (i.e., 1 g), while X_1 is held constant. Here, we may be tempted to conclude that the predictor variable X_1 is much more important than X_2. After observation, we realize that in fact X_2 is more important than X_1 simply because if X_2 changes by the same amount, that is, 1 kg, while X_1 is held constant, then the response variable will change by 700 units. It is therefore more meaningful to use a transformation on the response and the predictor variables so that the regression coefficients are dimensionless. These regression coefficients are usually referred to as *standardized regression coefficients.* This is achieved by using the transformation

$$U_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, k \tag{16.5.1}$$

where

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad S_j^2 = \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2, \quad S_y^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{16.5.2}$$

and S_j^2 is the corrected sum of squares for predictor X_j. It can easily be seen that with these transformations, the regression model (16.2.2) takes the form (see also (16.2.3))

$$Y_i = \gamma_0 + \gamma_1 U_{i1} + \gamma_2 U_{i2} + \cdots + \gamma_k U_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{16.5.3}$$

where

$$\gamma_0 = \beta_0 + \sum_{j=1}^{k} \beta_j \bar{X}_j, \quad \gamma_j = S_j \beta_j, \quad j = 1, 2, \ldots, k \tag{16.5.3a}$$

We remark that as is easily seen,

$$\sum_{i=1}^{n} U_{ij} = 0, \quad \sum_{i=1}^{n} U_{ij}^2 = 1, \qquad j = 1, 2, \ldots, k \tag{16.5.3b}$$

In matrix form, model (16.5.3) may be rewritten as

$$Y = U\gamma + \varepsilon \tag{16.5.3c}$$

where the $[(k+1) \times 1]$ vector γ is such that $\gamma' = (\gamma_0, \gamma_1, \ldots, \lambda_k)$, and where the matrix U $(n \times (k+1))$ is given by

$$U = \begin{pmatrix} 1 & U_{11} & U_{12} & \cdots & U_{1k} \\ 1 & U_{21} & U_{22} & \cdots & U_{2k} \\ 1 & U_{31} & U_{32} & \cdots & U_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & U_{n1} & U_{n2} & \cdots & U_{nk} \end{pmatrix} \tag{16.5.3d}$$

Now, using the properties of the U_{ij}s stated in (16.5.3b), we find that the $(k+1) \times (k+1)$ matrix $U'U$ takes the form

$$U'U = \left(\begin{array}{c|ccccc} n & 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ 0 & r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ 0 & r_{13} & r_{23} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{array}\right) \tag{16.5.3e}$$

where $r_{wv} = \sum_{i=1}^{n} U_{iw} U_{iv}$. It can be shown that $-1 \leq r_{st} \leq 1$, for $s \neq t$. We call model (16.5.3), or its equivalent (16.5.3c), the *standardized multiple linear regression model.* The least-squares estimates of the regression coefficients $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_k)'$ are given by

$$\hat{\gamma} = (U'U)^{-1} U'Y \tag{16.5.4}$$

The lower matrix $R_{(k \times k)}$ of the matrix $U'U$ is called the *sample correlation matrix* of the predictor variables X_j. The regression coefficients $\hat{\gamma}$ are called the *standardized regression coefficients.* Note here that even though predictor variables are not random variables, r_{ij} does provide information about the linear dependency between the predictor variables X_i and X_j.

16.5.1 Multicollinearity

So far, in our discussion on linear regression in Chapter 15 and in this chapter, various interests have been in (i) determining which predictor variables are valuable in predicting the behavior of the dependent variable, or which are not valuable in predicting the dependent variable and therefore should be dropped from the regression model, and (ii) looking for more predictor variables or additional terms that should be included in the model to improve significantly the predictability of the dependent variable. In the first case, we usually look at the p-values, while in the latter case, we look at the value of the adjusted coefficient of determination. However, another important aspect of a regression problem that we have not discussed is the effect of including a predictor variable in the presence of another predictor variable when the two variables are highly correlated. Including highly correlated predictor variables in a regression model is called *multicollinearity.*

Earlier, we noted that in dealing with the model

$$Y = X\beta + \varepsilon$$

the least-squares estimate of β is given by

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Obviously, these estimates are well defined if we assume that X is a full rank matrix (or, equivalently, $X'X$ is nonsingular). However, sometimes situations may arise in which $X'X$ is singular or nearly singular. If the vectors of the predictor variables X_i and X_j for $i \neq j$ are orthogonal, then $X'X$ is nonsingular. However, if the vectors of predictor variables are exactly linearly dependent, then $X'X$ is singular. If observation vectors on predictor variables are nearly linearly dependent, then the matrix $X'X$ is nearly singular. If exact or near-linear dependence exists among the predictor variables, then we say that there is *multicollinearity* in the model. The reader may adopt the following procedure to detect and eliminate multicollinearity.

Calculate the *simple correlation matrix* between the predictor variables. That is, compute the matrix R, where

$$R_{(k \times k)} = (r_{st}) \tag{16.5.5}$$

and where

$$r_{st} = \frac{\sum_{i=1}^{n}(X_{is} - \bar{X}_s)(X_{it} - \bar{X}_t)}{\sqrt{\sum_{i=1}^{n}(X_{is} - \bar{X}_s)^2 \sum_{i=1}^{n}(X_{it} - \bar{X}_t)^2}} \tag{16.5.6}$$

is the simple correlation between X_r and X_s. If r_{st} is large (i.e., r_{st}^2 is near 1), then there is a serious chance that there is multicollinearity in the model. One method of eliminating the multicollinearity is to drop one of the predictor variables, either X_i or X_j, from the model. Note that if $X'X$ still continues to be singular or nearly singular after dropping X_i or X_j, then this implies that some other predictor variables remaining in the new model are highly correlated. In such a case, repeat the above procedure and drop one of the remaining predictor variables. One can continue to repeat the above procedure until the multicollinearity disappears from the model.

16.5.2 Consequences of Multicollinearity

1. If some predictor variables are correlated, then the estimates of the regression coefficients for different samples may vary significantly. As a result the experimenter's ability to estimate the regression coefficients and interpret the data properly may be severely impaired. The usual interpretation of a regression coefficient, as the rate of change in the expected value of the dependent variable when a given predictor variable is increased by one unit, while all other predictor variables are held constant, may no longer be valid.
2. Due to collinearity among the predictor variables, variances among regression coefficients may be inflated. This inflation in turn causes the same types of problems as described in item 1. To illustrate, we consider two sets of two predictor variables, one with uncorrelated variables and the other with highly correlated predictor variables. We want to see how, in each case, the variances of regression coefficients are affected, assuming that we fit a standardized multiple linear regression model with two predictor variables.

The two data sets on predictor variables are

Data set 1: X_1 and X_2 are uncorrelated												
X_1	7	5	5	7	5	7	7	5	7	5	5	7
X_2	7	5	5	5	7	7	5	7	5	7	5	7

In this case, the correlation matrix R and its inverse are given by

$$R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad R^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Hence, the variances of standardized regression coefficients are given by (16.5.3e)–(16.5.4)

$$\frac{Var(\hat{\gamma}_1)}{\sigma^2} = \frac{Var(\hat{\gamma}_2)}{\sigma^2} = 1$$

Data set 2: X_1 and X_2 are highly correlated												
X_1	5.0	4.0	6.0	7.0	7.2	6.5	8.2	6.8	5.8	5.4	5.7	5.8
X_2	5.3	4.2	5.8	6.5	6.8	7.1	7.8	6.5	6.2	5.1	5.5	5.9

In this case, the correlation matrix R and its inverse are given by

$$R = \begin{pmatrix} 1 & 0.948 \\ 0.948 & 1 \end{pmatrix}, \quad R^{-1} = \begin{pmatrix} 9.87 & -9.36 \\ -9.36 & 9.87 \end{pmatrix}$$

Here the variances of standardized regression coefficients are then given by

$$\frac{Var(\hat{\gamma}_1)}{\sigma^2} = \frac{Var(\hat{\gamma}_2)}{\sigma^2} = 9.87$$

Obviously, in this case, variances of regression coefficients are inflated when there is multicollinearity. The diagonal elements in the R^{-1} matrix are called the *variance inflation factors*, usually denoted by *VIF*s. A *VIF* is a good measure of multicollinearity.

For further discussion on multicollinearity, the reader is referred to Kutner et al. (2004), Montgomery et al. (2006), and Myers (1990).

16.6 BUILDING REGRESSION TYPE PREDICTION MODELS

Once the influential predictor variables, $X_1, X_2, \ldots, X_k$, are identified, if the response or dependent variable is denoted by Y, then the problem is to find the function $f(X_1, X_2, \ldots, X_k)$ that is the best predictor for Y at the given values of $X_1, X_2, \ldots, X_k$. If the general form of f is known, then there is only the problem of fitting f to the data. This problem, in this chapter, has already been discussed for f a linear function in the regression parameters. We now consider the problem when f is unknown.

16.6.1 First Variable to Enter into the Model

Here, we want to find which of the possible predictor variables, $X_1, X_2, \ldots, X_k$, should be considered first. We can approach this problem in two ways. The first approach is to construct a model using all predictor variables $X_1, X_2, \ldots, X_k$, that are likely to affect the dependent variable Y. Then, we eliminate the predictor variables that do not make any significant contribution to predicting the dependent variable, Y. This procedure is called the *backward eliminating procedure.* The second approach is to consider the predictor variables $X_1, X_2, \ldots, X_k$ one by one. This procedure is called the *stepwise regression procedure*, sometimes the *forward selection procedure.* In this procedure, we first take the variable likely to be the most important or significant. Then we take the next one and construct a model including the variables already selected. Each time we test the significance of the effect of each new variable added into the model, with the variables that are already in the model. We proceed until a suitable model is constructed. Here "suitable" implies that we continue until the procedure is stopped according to some stopping criterion. We now consider this procedure.

This stepwise regression procedure proceeds by our computing the linear correlation between the values of Y and the corresponding values of X_t. We let $(Y_1, X_{1t}), (Y_2, X_{2t}), \ldots, (Y_n, X_{nt})$ be the points formed using the observations Y_i and the corresponding values X_{it} of the tth predictor variable. Then, the linear correlation between Y and X_t, denoted by r_{yt}, is given by

$$r_{yt} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_{it} - \bar{X}_t)}{\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sum_{i=1}^{n}(X_{it} - \bar{X}_t)^2\right]^{1/2}} \tag{16.6.1}$$

We calculate r_{yt} for $t = 1, 2, \ldots, k$ and select the X_t for which $|r_{yt}|$ is the largest. If r_{ym} is the largest in absolute value, then X_m is the first predictor variable to be considered. We now rename the predictor as X_1, say, and find the first-order linear regression model $\hat{Y} = f(X_1)$; that is, we fit Y to X_1, obtaining

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 \tag{16.6.2}$$

We check the significance of the coefficient of β_1 using the t statistic given by

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}, \quad \hat{\sigma}_{\hat{\beta}_1} = \sqrt{MSE \times a_{11}} \tag{16.6.3}$$

Also, we compute the coefficient of determination R^2. After considering X_1, the next step is to find the next variable to enter into the regression function. One way of doing this is to observe the residuals. The residuals, after removing the effect of X_1 from Y, are given by

$$U_{iy} = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1}) \tag{16.6.4}$$

Now, consider all other predictor variables $X_2, \ldots, X_k$. We wish to remove the effect of X_1 from all these variables. For example, with X_2 playing the role of the dependent variable, fit X_2 linearly to X_1 using the points $(X_{i1}, X_{i2}), \quad i = 1, 2, \ldots, n$. This results in the fitted model, say $\hat{X}_{i2} = \hat{c} + \hat{d}X_{i1}$. We now compute the residuals in each case

$$U_{i2} = X_{i2} - \hat{X}_{i2} = X_{i2} - (\hat{c} + \hat{d}X_{i1}), \quad i = 1, 2, \ldots, n \tag{16.6.5}$$

We repeat the process with $X_3, \ldots, X_k$ so that residuals $U_{i2}, U_{i3}, \ldots, U_{ik}$ are now available. We next find the $(k-1)$ linear correlations between U_{iy} of (16.6.4) and U_{it}, $t = 2, 3, \ldots, k$. These correlations are called the *partial correlations* because the effect of X_1 has been removed, or more accurately, they are the *partial correlations* between Y and X_t, $t = 2, \ldots, k$. We then select that X_i for which the partial correlation is a maximum in absolute value. It is to be the second variable we enter into the model, and we *let it be denoted* by X_2. Now we fit a regression of Y on X_1 and X_2. We denote this fitted regression equation by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \tag{16.6.5a}$$

Again, we compute the coefficient of determination R^2. We would then see that the R^2 value is improved, perhaps substantially by including the additional variable X_2. We continue in this way, adding one predictor variable at a time.

Each time the R^2 value will increase, but the increase at some stage will be insignificantly small. It is at this point that we stop the process. Each time we add a new variable, we test the significance of all the regression coefficients, including those that have already been tested. For example, after fitting the model (16.6.5a), we test the significance of the coefficients β_1 and β_2, which are individually tested by using the t-statistics defined in a similar fashion to (16.6.3).

The reader may wonder why one should again test for the significance of β_1 of model (16.6.5a) since the coefficient β_1 of model (16.6.2) has, in our overall procedure, been tested for significance. But apart from the fact that we are using generic notation, notice that the role of $\hat{\beta}_1$ in model (16.6.2) is very different than the role of $\hat{\beta}_1$ in model (16.6.5a), which is in the presence of $\hat{\beta}_2$. Indeed, the normal equations for (16.6.5a) show that the solution for $\hat{\beta}_1$ involves X_{i2} while, of course, the solution for $\hat{\beta}_1$ of normal equations for (16.6.2) does not. Thus, the β_1's are different, and it is important to test the significance of regression coefficients at each stage of testing.

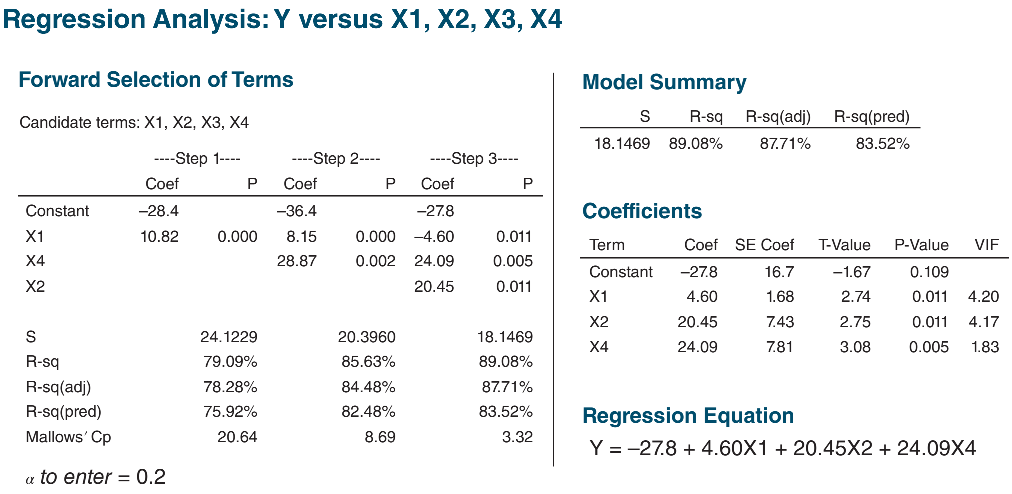
We continue the process after considering X_1 and X_2, by calculating the partial correlation between Y and $X_3, \ldots, X_k$ after removing the effect of X_1 and X_2. At this stage, we should also look at X_1X_2 because there is a possibility that X_1X_2 should be present in the regression model. Hence, we consider the partial correlation between Y and X_1X_2 after removing the effect of X_1 and X_2 from X_1X_2. Among all these variables (i.e., among variables $X_3, \ldots, X_k, X_{k+1} = X_1X_2$), the one with maximum partial correlation may be taken as the next variable to enter in the model. Again, we compute the t statistics with the new variables that have been included and calculate R^2. Then, we test the significance of each regression coefficient. If any coefficient is found insignificant at any stage, we drop the corresponding variable from the model. So at each stage, we drop all the variables previously dropped and then consider the remaining variables, together with the various products of the variables present in our built-up model, and look for the next variable to add to the regression model. This process is continued until there is no significant improvement in R^2. The resulting model is a prediction model for Y.

Example 16.6.1 (Home prices data) *Refer to the data in Example 16.4.2 (see Table 16.4.3). Fit a multiple linear regression model using the method of stepwise regression (forward selection) in MINITAB and R.*

Solution:
MINITAB

To fit a regression model using stepwise regression in MINITAB, we proceed as follows:

1. Enter the data in columns C1–C5.
2. Select **Stat** > **Regression** > **Regression** > **Fit Regression Model.** A dialog box "Regression" appears. In this dialog box enter Y in the box for **Response** and X_1, X_2, X_3, and X_4 in box for **Continuous predictors**. Then select **Stepwise** from this dialog box. A new dialog box appears. In this new dialog box, select from the pull down menu **Method: Forward selection** from this new box. Also, enter the desired value of alpha (α) value in the box next to **Alpha to enter:** and from the pull down menu **Display the table of model selection details** select **Include details for each step**. MINITAB uses the alpha value as the criterion for adding a variable to the model. All the variables with p-value less than the alpha value are added to the model. Click **OK** in each dialog box. Then the MINITAB output appears in the Session window as shown here.



```
Regression Analysis: Y versus X1, X2, X3, X4

Forward Selection of Terms                        Model Summary

Candidate terms: X1, X2, X3, X4                         S    R-sq  R-sq(adj)  R-sq(pred)
                                                  18.1469  89.08%     87.71%      83.52%
               ----Step 1----  ----Step 2----  ----Step 3----
               Coef       P    Coef       P    Coef       P
                                                  Coefficients
Constant      –28.4           –36.4           –27.8
X1            10.82   0.000    8.15   0.000   –4.60   0.011   Term       Coef  SE Coef  T-Value  P-Value   VIF
X4                            28.87   0.002   24.09   0.005   Constant  –27.8     16.7    –1.67    0.109
X2                                            20.45   0.011   X1         4.60     1.68     2.74    0.011  4.20
                                                              X2        20.45     7.43     2.75    0.011  4.17
S                   24.1229         20.3960         18.1469   X4        24.09     7.81     3.08    0.005  1.83
R-sq                 79.09%          85.63%          89.08%
R-sq(adj)            78.28%          84.48%          87.71%
R-sq(pred)           75.92%          82.48%          83.52%   Regression Equation
Mallows′ Cp           20.64            8.69            3.32
                                                              Y = –27.8 + 4.60X1 + 20.45X2 + 24.09X4
α to enter = 0.2
```

This forward selection procedure that uses the stepwise regression method takes three steps to fit the model. The first step determines a constant term and the first variable, which in this case is X_1. In the second and the third steps it adds X_4 and X_2, respectively, to the model. The final fitted model is

$$\hat{Y} = -27.8 + 4.60X_1 + 20.45X_2 + 24.09X_4$$

Notice that in this fitted model, the predictor variable X_3 is absent. This could be because its observed level of significance is high or its contribution to the value of R^2 is negligible, or both.

Notes:

1. In each step, the values of the constant term and the other regression coefficients determined previously change.
2. The value "Alpha-to-Enter" is kept fairly high so that no predictor variable is thrown out prematurely.
3. The output includes values of various statistics that are useful in model selection (see Example 16.7.1).

USING R

To obtain the stepwise regression model, we use 'stepAIC()' function in R MASS package. It chooses the best model by Akaike Information Criteria (AIC) (see Akaike, 1973). The additional option 'direction' can be used to specify the selection criteria. That is, option = "both" applies both forward and backward selection, option = "backward" applies backward selection, and option = "forward" applies forward selection. Finally, it returns the best selected model.

```
library(MASS)
Y = c(160,136,189,116,136,255,209,135,265,115,114,219,299,200,190,
209,200,150,118,135,230,155,118,159,199,200,260,210)
X1 = c(19,17,22,12,19,24,18,15,28,12,18,21,26,21,22,24,20,18,12,17,
22,18,12,17,21,23,24,21)
X2 = c(3,3,4,2,4,5,4,2,4,2,2,4,5,4,4,4,4,3,2,3,4,3,2,3,3,4,5,4)
X3 = c(2,2,2,1,2,3,2,2,3,1,2,3,3,2,2,2,2,2,1,2,3,2,1,2,2,2,3,2)
X4 = c(2,2,2,1,1,3,3,1,3,1,2,2,3,2,2,2,2,2,2,2,3,2,2,2,2,2,3,2)


#Fitting MLR model using predictors X1, X2, X3, and X4
model = lm(Y ∼ X1 + X2 + X3 + X4)
step.model = stepAIC(model, direction = "both")
summary(step.model)
```

#R summary output

	Estimate	Std. Error	t value	Pr(>$\|t\|$)
(Intercept)	−27.8426	16.7115	−1.67	0.1087
X1	4.6037	1.6820	2.74	0.0115
X2	20.4535	7.4286	2.75	0.0111
X4	24.0901	7.8140	3.08	0.0051

Note that the summary results are identical to those obtained by using MINITAB.

16.7 RESIDUAL ANALYSIS AND CERTAIN CRITERIA FOR MODEL SELECTION

16.7.1 Residual Analysis

Consider the multiple linear regression model

$$Y = X\beta + \varepsilon \tag{16.7.1}$$

If the covariance matrix is of the form $Cov(\varepsilon) = D$, where D is a diagonal matrix whose diagonal elements are not equal, we say that there is *heteroskedasticity* in the model in the sense that the variances of ε_is are not all equal, but the ε_i's are uncorrelated. That is, the assumption of equal variances is violated. In the case of multiple linear regression, as in simple linear regression, plotting the residuals $Y_i - \hat{Y}_i$ against the fitted values $\hat{Y}_i$ is usually helpful to check whether heteroskedasticity is present. If all the scattered points fall within a horizontal band, as shown in Figure 16.7.1, then this indicates the absence of heteroskedasticity: in other words, the assumption of equality of variances of the ε_i's is satisfied.

If, however, the plot of residuals versus fitted values forms a pattern, such as a parabola or a funnel shape, then heteroskedasticity is present and consequently our assumption of equal variances is violated. The residual plots for the salary data in Example 16.4.1 are as shown in Figure 16.7.2a–d.

Note that these plots show no significant abnormalities about the model except the data seems to be skewed to the right.

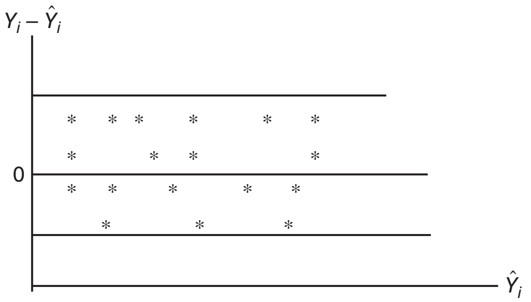


Figure 16.7.1 Desirable residual plot.

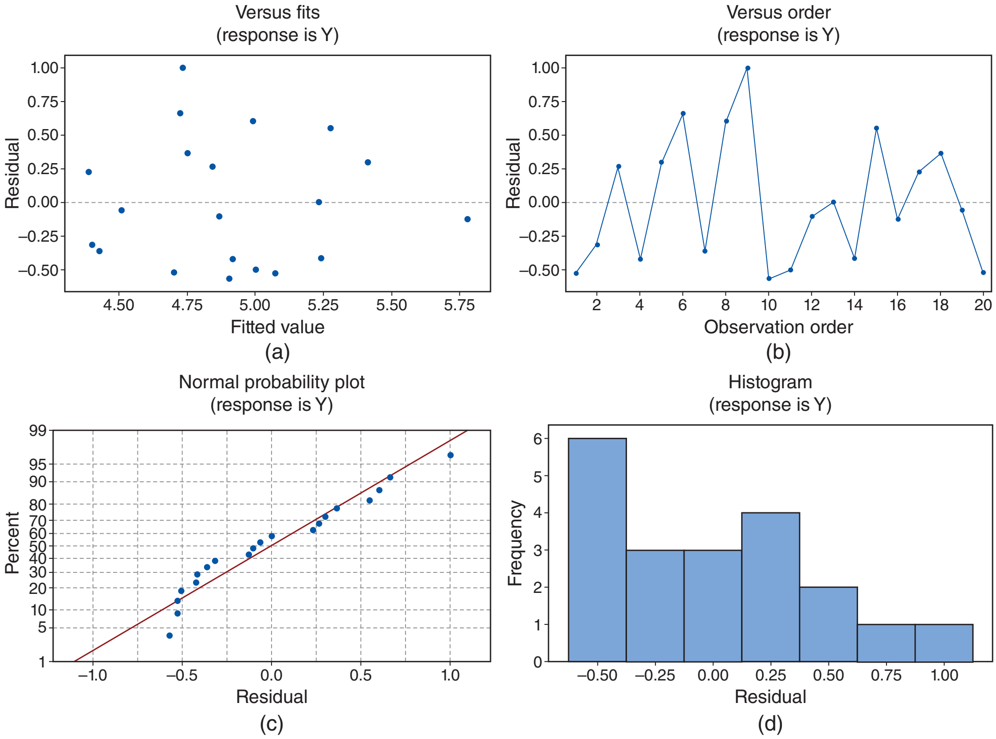


Figure 16.7.2 Residual plots: (a) residuals versus fitted values, (b) residuals versus observation order, (c) normal probability plot of residuals, and (d) histogram of residuals.

16.7.2 Certain Criteria for Model Selection

Suppose that we want to build a multiple linear regression using k possible predictor variables. Then we have 2^k possible selections of subsets of k predictor variables. For example, if the predictor variables under consideration are $\{X_1, X_2, X_3\}$, then $2^3 = 8$ possible subsets are

$$\{\phi\}, \{X_1\}, \{X_2\}, \{X_3\}, \{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}, \quad \text{and} \quad \{X_1, X_2, X_3\}$$

There are various criteria for selecting a subset regression model, but we will discuss only four of them, namely R^2, R^2_{adj}, C_p, and PRESS.

Coefficient of Multiple Determination—$\mathbf{R^2}$

The R^2 *coefficient of multiple determination* is the most commonly used criterion for a good model. The coefficient of multiple determination is defined as

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = \frac{SSR}{SS_{Total}} \tag{16.7.2}$$

Alternatively, R^2 can also be defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = 1 - \frac{SSE}{SS_{Total}} \tag{16.7.3}$$

From (16.7.2) we can see that R^2 measures the proportion of variation in the response explained by regression, that is, by the presence of the k predictor variables $X_1, \ldots, X_k$. Note further that

$$0 \leq R^2 \leq 1 \tag{16.7.4}$$

and from (16.7.3) it follows that $R^2 = 1$ implies that $SSE = 0$, in which case all the (observed) errors are zero; that is, $R^2 = 1$, if and only if there is a perfect fit. The value of R^2 always increases as an additional predictor variable is included in the model. Obviously, R^2 is largest when all the predictor variables under consideration are included in the model. In practice, the addition of any new predictor variable in the model is not considered useful when such an addition increases the value of R^2 by an insignificantly small amount.

Adjusted Coefficient of Multiple Determination—R^2_{adj}

As noted above, one of the drawbacks of R^2 is that its value always increases when any new predictor variable is included in the model. Hence, an alternative criterion that is commonly used is the *adjusted coefficient of multiple determination*, which is defined as

$$R^2_{adj} = 1 - \frac{n-1}{n-r-1}\frac{SSE}{SS_{Total}} = 1 - (n-1)\frac{MSE}{SS_{Total}} \tag{16.7.5}$$

where r is the number of parameters (not including β_0) in the model; that is, the number of predictor variables in the model. From (16.7.5), we easily see that R^2_{adj} decreases if and only if MSE increases, since $(n-1)/SS_{Total}$ is fixed. Further, if, with the addition of more factors X_i into the model, we find that SSE does not decrease significantly, then MSE will increase since as a small decrease in SSE may not offset the loss due to the corresponding error degrees of freedom $(n - r - 1)$. Hence, R^2_{adj} will consequently decrease.

Mallows' C_p Statistic

This criterion is related to the total *expected mean squared error* of the n-fitted values for each subset regression model. The mean square error of $\hat{Y}_i$ at a given value X_i of X using any specific subset regression model is given by

$$E(\hat{Y}_i - E(Y_i))^2 \tag{16.7.6}$$

where $E(Y_i)$ is the true mean response at $X = (X_{i1}, X_{i2}, \ldots, X_{ip})$ and p is the number of predictor variables in the model. Now, (16.7.6) can be written as

$$\begin{aligned} E(\hat{Y}_i - E(Y_i))^2 &= E(\hat{Y}_i - E(\hat{Y}_i) + E(\hat{Y}_i) - E(Y_i))^2 \\ &= E(\hat{Y}_i - E(\hat{Y}_i))^2 + (E(\hat{Y}_i) - E(Y_i))^2 \\ &= Var(\hat{Y}_i) + (E(\hat{Y}_i) - E(Y_i))^2 \end{aligned} \tag{16.7.7}$$

where $E(\hat{Y}_i) - E(Y_i)$ is the bias at the data point $X = (X_{i1}, X_{i2}, \ldots, X_{ip})$ and p is the number of predictor variables in the model. From (16.7.7), we see that the total mean squared error for all the n-fitted values is given by

$$\sum_{i=1}^{n} E(\hat{Y}_i - E(Y_i))^2 = \sum_{i=1}^{n} Var(\hat{Y}_i) + \sum_{i=1}^{n} [E(\hat{Y}_i) - E(Y_i)]^2 \tag{16.7.8}$$

We now define the standardized total mean squared error as

$$\begin{aligned} \Gamma_p &= \tfrac{1}{\sigma^2} \sum_{i=1}^{n} E(\hat{Y}_i - E(Y_i))^2 \\ &= \tfrac{1}{\sigma^2} \left(\sum_{i=1}^{n} Var(\hat{Y}_i) + \sum_{i=1}^{n} [E(\hat{Y}_i) - E(Y_i)]^2 \right) \end{aligned} \tag{16.7.9}$$

It can be shown that

$$\sum_{i=1}^{n} Var(\hat{Y}_i) = p\sigma^2 \tag{16.7.10}$$

and that the expected value of the residual sum of squares for a model with p parameters (including β_0), that is, the number of predictor variables entered in the model + 1 for β_0, is given by

$$E(SSE_p) = (n-p)\sigma^2 + \sum_{i=1}^{n} [E(\hat{Y}_i) - E(Y_i)]^2 \tag{16.7.10a}$$

Now, substituting for $\sum_{i=1}^{n} Var(\hat{Y}_i)$ and $\sum_{i=1}^{n} [E(\hat{Y}_i) - E(Y_i)]^2$ from Equations (16.7.10) and (16.7.10a) in Equation (16.7.9), we obtain

$$\Gamma_p = \frac{E(SSE_p)}{\sigma^2} - (n - 2p) \tag{16.7.11}$$

Next, replacing $E(SSE_p)$ with the observed SSE_p and σ^2 with its estimator $\hat{\sigma}^2$, we obtain an estimator of Γ_p denoted by C_p:

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p) \tag{16.7.12}$$

If the model with p parameters does not have any lack of fit, then it can be shown that

$$\Gamma_p = p \tag{16.7.13}$$

because $E(SSE_p) = (n-p)\sigma^2$. Hence, models with no bias will produce a value of C_p that is expected to have value p, approximately. Thus, using C_p as a criterion for selecting an adequate model, we seek the value of C_p that is fairly close to p and the smallest among values of C_p as p varies. (Recall that p is the number of terms in the model, including the constant term.) Generally, we can say that a small value of C_p indicates that the model has a small total mean squared error, and that when C_p is also close to p, the bias of the regression model due to sampling error is also small. Models with C_p significantly greater than p are said to have large bias.

PRESS Statistic

An individual *PRESS* is defined as

$$e_{i(i)}^2 = (Y_i - \hat{Y}_{i(i)})^2 \tag{16.7.14}$$

where $\hat{Y}_{i(i)}$ is the predicted value of the *i*th observation obtained by deleting the *i*th observed response from the sample data points and estimating the fit of the model based on the remaining $(n-1)$ observations. The PRESS statistic identifies observations that have strong influence on the future prediction and is defined as

$$\text{PRESS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 \tag{16.7.15}$$

where h_{ii} is the *i*th diagonal element of the *HAT* matrix

$$H = X(X'X)^{-1}X' \tag{16.7.16}$$

and $e_i = y_i - \hat{y}_i$. (For more details on the PRESS statistic, see Montgomery et al., 2006.) A model with a small value of PRESS is considered to be a good fit to the data.

The PRESS statistic can be used to obtain another statistic, denoted by R_{pred}^2, which measures the capability of the model to predict future observations and is defined as

$$R_{pred}^2 = 1 - \frac{PRESS}{SS_{Total}} \tag{16.7.17}$$

We illustrate the foregoing measures in the following example.

Example 16.7.1 *Refer to Example 16.6.1.*

Step	1	2	3
R-Sq	79.09	85.63	89.08
R-Sq(adj)	78.28	84.48	87.71
Mallows Cp	20.60	8.70	3.30
PRESS	17,418.30	12,674.40	11,923.20
R-Sq(pred)	75.92	82.48	83.52

Note that when using MINITAB, the stepwise regression method automatically displays R^2, R_{adj}^2, and Mallows' C_p statistics as long as we select **Include details for each step** from the **Stepwise** window. However, to get the PRESS value, select **Results** from the **Regression** dialog box, a new dialog box appears. In this dialog box, select from the pull down menu **Display of Results:** select **Expanded tables**. Using the criteria discussed earlier in this section, the model obtained after Step 3 is the best among the three models obtained after Step 1, Step 2, and Step 3.

For further discussion of multiple linear regression, the reader is referred to Draper and Smith (1981), Freund and Wilson (1998), Kutner et al. (2004), Montgomery et al. (2006), and Myers (1990).

PRACTICE PROBLEMS FOR SECTIONS 16.6 AND 16.7

1. The quality of pinot noir wine is believed to be related to clarity (X_1), aroma (X_2), body (X_3), flavor (X_4), oakiness (X_5), and region (a qualitative variable with three levels). The data for 38 wines is given below (data from MINITAB data files).

X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
1.0	3.3	2.8	3.1	4.1	0	1	9.8	0.9	3.4	5.0	3.4	3.4	1	0	7.9
1.0	4.4	4.9	3.5	3.9	0	1	12.6	0.9	6.4	5.4	6.6	4.8	0	0	15.1
1.0	3.9	5.3	4.8	4.7	0	1	11.9	1.0	5.5	5.3	5.3	3.8	0	0	13.5
1.0	3.9	2.6	3.1	3.6	0	1	11.1	0.7	4.7	4.1	5.0	3.7	1	0	10.8
1.0	5.6	5.1	5.5	5.1	0	1	13.3	0.7	4.1	4.0	4.1	4.0	1	0	9.5
1.0	4.6	4.7	5.0	4.1	0	1	12.8	1.0	6.0	5.4	5.7	4.7	0	0	12.7
1.0	4.8	4.8	4.8	3.3	0	1	12.8	1.0	4.3	4.6	4.7	4.9	1	0	11.6
1.0	5.3	4.5	4.3	5.2	0	1	12.0	1.0	3.9	4.0	5.1	5.1	0	1	11.7
1.0	4.3	4.3	3.9	2.9	0	0	13.6	1.0	5.1	4.9	5.0	5.1	1	0	11.9
1.0	4.3	3.9	4.7	3.9	0	1	13.9	1.0	3.9	4.4	5.0	4.4	1	0	10.8
1.0	5.1	4.3	4.5	3.6	0	0	14.4	1.0	4.5	3.7	2.9	3.9	1	0	8.5
0.5	3.3	5.4	4.3	3.6	1	0	12.3	1.0	5.2	4.3	5.0	6.0	1	0	10.7
0.8	5.9	5.7	7.0	4.1	0	0	16.1	0.8	4.2	3.8	3.0	4.7	0	1	9.1
0.7	7.7	6.6	6.7	3.7	0	0	16.1	1.0	3.3	3.5	4.3	4.5	0	1	12.1
1.0	7.1	4.4	5.8	4.1	0	0	15.5	1.0	6.8	5.0	6.0	5.2	0	0	14.9
0.9	5.5	5.6	5.6	4.4	0	0	15.5	0.8	5.0	5.7	5.5	4.8	0	1	13.5
1.0	6.3	5.4	4.8	4.6	0	0	13.8	0.8	3.5	4.7	4.2	3.3	0	1	12.2
1.0	5.0	5.5	5.5	4.1	0	0	13.8	0.8	4.3	5.5	3.5	5.8	0	1	10.3
1.0	4.6	4.1	4.3	3.1	0	1	11.3	0.8	5.2	4.8	5.7	3.5	0	1	13.2

Region 1 is denoted by assigning $(X_6, X_7) = (1, 0)$, region 2 by $(X_6, X_7) = (0, 1)$, and region 3 by $(X_6, X_7) = (0, 0)$.

(a) Apply the stepwise regression method to fit the following model to these data:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

Use $\alpha = 0.15$ to enter predictor variables and $\alpha = 0.15$ to remove them.

(b) Use the PRESS or R^2_{pred} criterion to find a suitable regression model.

2. Refer to Problem 1. Use Mallows' C_p criterion to see if the model chosen is the same model chosen in Problem 1.

3. Refer to Problem 1. When applying the stepwise method, always include the predictor variables X_1, X_2. Compute the Mallows' C_p, PRESS, and R^2_{pred} statistics. Find a suitable regression model using Mallows' C_p criterion and the PRESS or R^2_{pred} criterion. Compare the models chosen here and in Problems 1 and 2, and comment.

4. (a) Fit the following model to the data in Problem 1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_3 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

   (b) Determine by using the PRESS or R^2_{pred} criterion if this new model is a better fit.

5. Refer to the data on the observed mole fraction solubility in Problem 19 of Review Practice Problems.

   (a) Fit the following model using the stepwise regression method

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \varepsilon$$

   (b) Compute Mallows' C_p statistic and comment on its value.

6. Refer to Problem 7 of Section 16.3:

   (a) Apply the stepwise regression method to fit the following model to the data in Problem 7 of Section 16.3.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

   Use $\alpha = 0.15$ to enter predictor variables and $\alpha = 0.15$ for possible removal.

   (b) Use the PRESS or R^2_{pred} criterion to evaluate the regression model.

7. (a) Fit the following model using the stepwise regression method to the data in Problem 7 of Section 16.3:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_1 X_5 \\ + \beta_8 X_5 X_6 + \varepsilon$$

   (b) Determine using the PRESS or R^2_{pred} criterion whether this new model yields a better fit than the fit to the model of Problem 6 above.

8. Apply the stepwise regression method to fit the following model to the data in Problem 7 of Section 16.3:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_3^2 + \beta_8 X_4^2 \\ + \beta_9 X_3 X_4 + \varepsilon$$

   Use $\alpha = 0.15$ to enter predictor variables and $\alpha = 0.15$ to remove, and use the PRESS or R^2_{pred} criterion to evaluate the regression model.

16.8 LOGISTIC REGRESSION

Logistic regression is commonly used when the response variable (i.e., the dependent variable) is a binary variable. For example, the response of interest may center on whether a manufacturing company gets or does not get a contract, or a patient responds to a treatment or does not respond, or a smoker develops lung cancer or does not develop lung cancer. If a response is of the foregoing nature, then the response variable is a binary variable that takes the values 1 and 0 according to whether it has or does not have a certain characteristic. Here the response variable Y can be considered as a Bernoulli variable, and we let

$$P(Y = 1) = \theta, \quad P(Y = 0) = 1 - \theta \tag{16.8.1}$$

so that

$$E(Y) = \theta \tag{16.8.2}$$

A simple *logistic regression model* that is used frequently is defined by

$$\theta = E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \tag{16.8.3}$$

where X is a predictor variable and Y is the response or dependent variable. In Chapter 15, we learned that in the simple linear model, the expected value of the dependent variable Y can take any value on the real line; however, in the logistic regression model, the expected value of Y takes a value between 0 and 1. Hence, when Y is a binary (Bernoulli) variable, it cannot be modeled with the usual simple linear regression model. Nevertheless, the logistic regression model in (16.8.3) can be expressed in a simple form by using a transformation called the *logit transformation* defined by

$$\eta = \ln\left(\frac{\theta}{1-\theta}\right) \tag{16.8.4}$$

so that

$$\eta = \beta_0 + \beta_1 X \tag{16.8.5}$$

which is a linear function of the predictor variable X. The expression $\theta/(1-\theta)$ in (16.8.4) is called the odds ratio. Note from (16.8.5) that the logit mean response η can take any value on the real line as the predictor variable X takes values between $-\infty$ and ∞.

Note further that if we have more than one predictor variables, then the logistic regression model may be stated as

$$\theta = E(Y) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)} \tag{16.8.6}$$

and by using the logit transformation defined in (16.8.4), we obtain

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{16.8.7}$$

We saw above that the response variable Y is a Bernoulli variable. Hence, the joint probability function of observations $Y_1, Y_2, \ldots, Y_n$ that are independent is given by

$$h(y_1, \ldots, y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \tag{16.8.8}$$

Taking the logarithm with base e of both sides, we obtain

$$\ln h(y_1, \ldots, y_n) = \sum_{i=1}^{n} y_i \ln\left(\frac{\theta}{1-\theta}\right) + \sum_{i=1}^{n} \ln(1-\theta) \tag{16.8.9}$$

Now, using Equations (16.8.3) through (16.8.5), we can express Equation (16.8.9) as

$$\ln h(y_1, \ldots, y_n) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \ln(1 + \exp(\beta_0 + \beta_1 X_i)), \tag{16.8.10}$$

which is the log-likelihood function of the unknown parameters β_0 and β_1. Thus, the (MLE) estimators b_0 and b_1 of β_0 and β_1, respectively, are obtained by maximizing the likelihood function in (16.8.10). Hence, we obtain, for a given value of X, the MLE $\hat{\theta}$ of θ as

$$\hat{\theta} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)} = \hat{\theta}(X) \tag{16.8.11}$$

and the MLE of $\eta = \ln[\theta/(1-\theta)]$, for given X, as

$$\hat{\eta} = \ln\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) \quad \text{with} \quad \hat{\theta} = \hat{\theta}(X) \tag{16.8.12}$$

given in (16.8.11).

In logistic regression, however, we do not have the same restrictive assumptions as in simple linear regression model for a number of reasons:

1. The relationship between the dependent and the predictor variables does not have to be linear.
2. The normality assumption for the dependent variable is not necessary.
3. The dependent variable does not have to be homoscedastic; that is, the assumption of homogeneity of variance does not need to hold.
4. The equations for determining b_0 and b_1 are obtained in the usual way, that is, by differentiating the log likelihood (16.8.10) with respect to β_0 and β_1 and setting the derivatives equal to zero, obtaining two equations whose solution is denoted by (b_0, b_1). However, the equations are nonlinear in (b_0, b_1), so (b_0, b_1) cannot be expressed in neat closed form. Indeed, the solutions require the use of iterative numerical analysis, but fortunately, MINITAB and other statistical packages carry this out automatically. See Example 16.8.1 below, where this is illustrated. Similar remarks hold for the case (16.8.6).
5. In logistic regression, interest centers on the estimation of the log-odds that the dependent variable Y takes the value 1, when the dependent variable corresponds to the characteristic of interest. For example, if we are interested in whether an exposed person gets a disease, we estimate the log-odds that a person who has been exposed will get the disease.

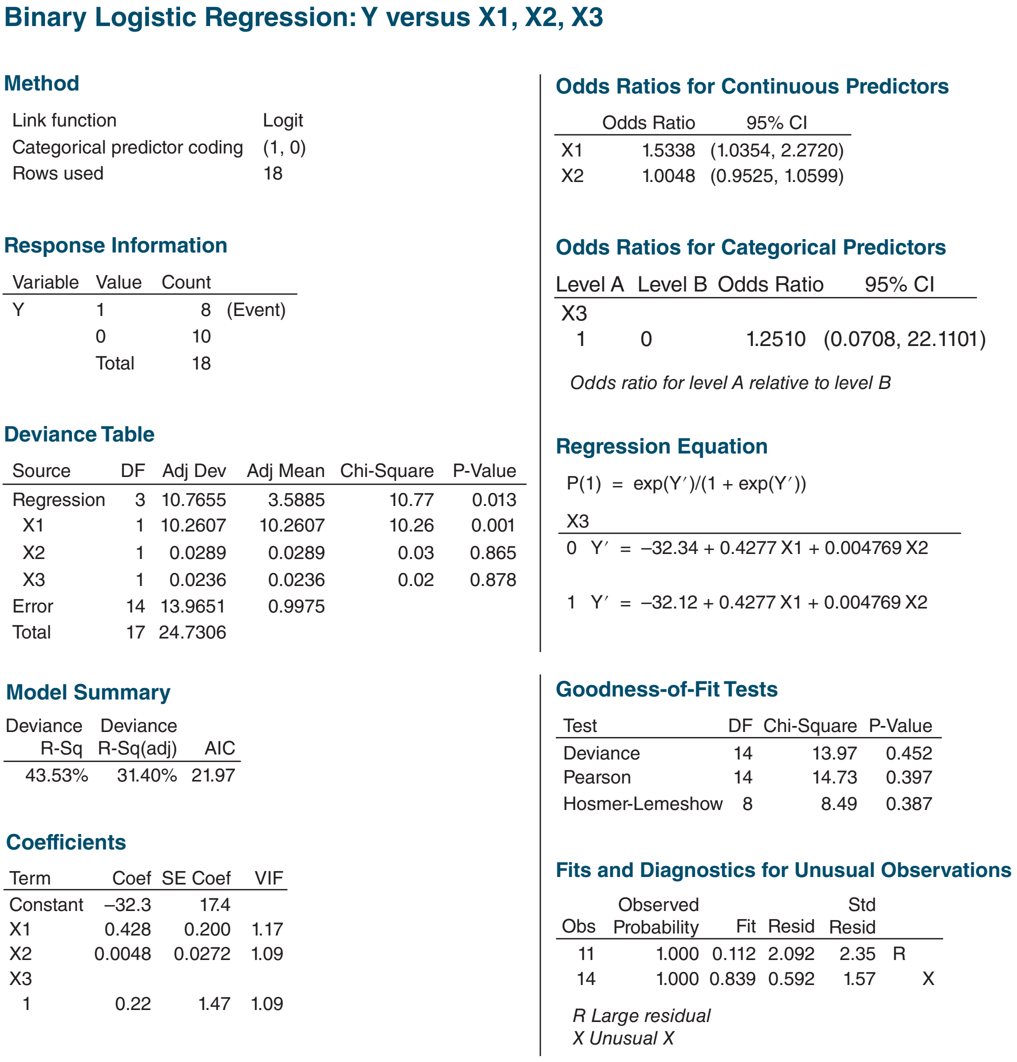
Example 16.8.1 (Obstructive coronary artery disease-related data) *Eighteen subjects (eight women and 10 men) with a history of high cholesterol levels are tested for obstructive coronary artery disease (OCAD). The results of the test, represented by a binary variable* Y *($Y = 1$ when a person has OCAD and $Y = 0$ when a person does not have OCAD) and the predictor variables X_1 (age), X_2 (total cholesterol level), and X_3 (categorical predictor variable, 1 for women and 0 for men) are shown in Table 16.8.1. Analyze these data using logistic regression.*

Solution:
MINITAB

Enter the data in the MINITAB worksheet. Then from the Menu bar select **Stat** > **Regression** > **Binary Logistic Regression** > **Fit Binary Logistic Regression Model**. Under **Response/frequency format response** enter Y in the box next to **Response:**, the continuous predictor variables in the box below **Continuous predictors:**, and the categorical predictor variable in the box below **Categorical predictors:**. From Graphs, Options, Results, and Storage select the desired entries and click **OK**. The MINITAB printout is shown here.

Table 16.8.1 Results of the OCAD test.

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
0	69	247	1	0	67	212	1
1	77	238	0	1	68	249	0
1	73	229	0	0	74	234	0
0	74	241	0	1	74	233	1
1	79	245	1	1	78	133	0
0	66	231	0	0	73	218	0
0	62	236	1	0	69	219	1
1	80	235	0	1	78	216	1
0	62	234	0	0	65	210	1



```
Binary Logistic Regression: Y versus X1, X2, X3

Method
 Link function                    Logit
 Categorical predictor coding     (1, 0)
 Rows used                        18

Response Information
 Variable  Value   Count
 Y         1           8  (Event)
           0          10
           Total      18

Deviance Table
 Source       DF  Adj Dev   Adj Mean  Chi-Square  P-Value
 Regression    3  10.7655     3.5885       10.77    0.013
   X1          1  10.2607    10.2607       10.26    0.001
   X2          1   0.0289     0.0289        0.03    0.865
   X3          1   0.0236     0.0236        0.02    0.878
 Error        14  13.9651     0.9975
 Total        17  24.7306

Model Summary
Deviance   Deviance
    R-Sq  R-Sq(adj)    AIC
  43.53%     31.40%  21.97

Coefficients
Term        Coef  SE Coef   VIF
Constant   –32.3     17.4
X1         0.428    0.200  1.17
X2        0.0048   0.0272  1.09
X3
  1         0.22     1.47  1.09

Odds Ratios for Continuous Predictors
       Odds Ratio      95% CI
X1         1.5338   (1.0354, 2.2720)
X2         1.0048   (0.9525, 1.0599)

Odds Ratios for Categorical Predictors
Level A  Level B  Odds Ratio       95% CI
X3
  1        0            1.2510   (0.0708, 22.1101)

  Odds ratio for level A relative to level B

Regression Equation
 P(1)  =  exp(Y′)/(1 + exp(Y′))

 X3
 0   Y′  =  –32.34 + 0.4277 X1 + 0.004769 X2

 1   Y′  =  –32.12 + 0.4277 X1 + 0.004769 X2

Goodness-of-Fit Tests
 Test               DF  Chi-Square  P-Value
 Deviance           14       13.97    0.452
 Pearson            14       14.73    0.397
 Hosmer-Lemeshow     8        8.49    0.387

Fits and Diagnostics for Unusual Observations
       Observed                       Std
 Obs  Probability    Fit   Resid    Resid
  11        1.000  0.112   2.092     2.35  R
  14        1.000  0.839   0.592     1.57      X

  R Large residual
  X Unusual X
```

Interpretation of the **MINITAB** *Output* From the logistic regression table we see that X_2 and X_3 are not significant, but the *overall model is good* since the p-value is small (0.013). Thus we use

$$\hat{\eta} = -32.3432 + 0.4277X_1 + 0.0048X_2 + 0.2239X_3$$

That is,

$$\begin{aligned}\hat{\eta}(X_1, X_2, X_3) &= -32.3432 + 0.4277X_1 + 0.0048X_2 + 0.2239X_3 \\ \hat{\eta}(X_1 + 1, X_2, X_3) &= -32.3432 + 0.4277(X_1 + 1) + 0.0048X_2 + 0.2239X_3\end{aligned}$$

By taking the difference of the last two expressions, we obtain

$$\hat{\eta}(X_1 + 1, X_2, X_3) - \hat{\eta}(X_1, X_2, X_3) = 0.4277$$

From (16.8.4) we know that $\hat{\eta}$ is the estimate of log-odds, at (X_1, X_2, X_3), which implies that

$$\begin{aligned}&\log(odds(X_1 + 1, X_2, X_3)) - \log(odds(X_1, X_2, X_3)) \\ &\qquad = \log((odds(X_1 + 1, X_2, X_3))/(odds(X_1, X_2, X_3))) = 0.4277\end{aligned}$$

that is,

$$odds(X_1 + 1, X_2, X_3)/odds(X_1, X_2, X_3) = e^{0.4277} = 1.534$$

This can be interpreted as the estimated increase in probability of a person getting OCAD with a one-year increase in age, is 53%. In general, we can interpret this as saying that the estimated increase in probability of success with one unit increase in predictor variable X_1 is $(e^{\hat{\beta}_i} - 1)100\%$.

USING R

To fit logistic regression models in R, we can use the ‘glm()’ function. To perform the required analysis, we can run the following R-code.

```
Y = c(0,1,1,0,1,0,0,1,0,0,1,0,1,1,0,0,1,0)
X1 = c(69,77,73,74,79,66,62,80,62,67,68,74,74,78,73,69,78,65)
X2 = c(247,238,229,241,245,231,236,235,234,212,249,234,233,133,218,219,216,210)
X3 = c(1,0,0,0,1,0,1,0,0,1,0,0,1,0,0,1,1,1)
model = glm(Y ∼ X1 + X2 + X3, family = binomial)
anova(model, test =“Chisq”)
summary(model)

confint(model); # 95% CI for the coefficients
exp(coef(model)); # exponentiated coefficients
exp(confint(model)); # 95% CI for exponentiated coefficients
predict(model, type =“response”); # predicted probability of success
```

PRACTICE PROBLEMS FOR SECTION 16.8

1. The following data give MCAT scores X of 10 applicants who apply for admission to medical school. The result of each applicant is either accepted ($Y = 1$) or not accepted ($Y = 0$). Fit a logistic regression model and interpret your results.

Applicant	1	2	3	4	5	6	7	8	9	10
X	28	32	29	33	27	31	30	34	32	29
Y	0	1	0	1	0	0	0	1	0	1

2. A travel agency conducted a study to investigate the relationship between middle-class families' household income in thousands of dollars X and traveling at least 500 miles for vacation. Ten middle-class families were randomly selected and their family income and status ($Y = 1$ implies traveled and $Y = 0$ implies not traveled) of traveling at least 500 miles for their vacation were determined. The data obtained are given below. Fit a logistic regression model and interpret your results.

Family	1	2	3	4	5	6	7	8	9	10
X	83.0	82.3	72.8	78.0	82.3	71.8	87.3	82.3	86.1	70.0
Y	1	1	0	1	1	0	1	0	1	0

3. A manager of a manufacturing company conducted a study to investigate the relationship between the years X of service of an engineer and his/her performance ($Y = 1$ indicating satisfactory and $Y = 0$ not satisfactory). The following data give the results for 15 randomly selected engineers who work for that company. Fit a logistic regression model and interpret your results.

Engineer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	8	18	19	14	15	9	11	15	15	17	17	14	15	13	20
Y	1	0	1	0	1	0	1	0	1	1	1	0	0	0	1

16.9 CASE STUDIES

Case Study 1: (*Seal strength in surgical sponges*)[1] A paper package for gauze surgical sponges is created by feeding two rolls of paper into a machine to form the top and bottom

[1] Source: Based on data provided by Dr. Mary McShane, Southern Polytechnic State University, GA.

of the package. The top roll comes pretreated with adhesive applied in a chevron pattern. The seal is created when a heated platen is applied to the package for a certain length of time. Seal strength, measured in pounds per inch, is the primary response of interest.

An engineer wants to predict seal strength, given the machine settings of platen temperature, application pressure, and dwell time. In addition other factors may affect the strength of the seal. For example, she suspects that the relative humidity in the plant affects the adhesion. Two vendors are used to supply the paper-packaging material, and there may be a difference between their products. One hundred observations taken from production records from the previous five months are shown below. The data for this case study are available on the book website: www.wiley.com/college/gupta/statistics2e.

The data involves Platen temperature (°F) X_1, Application pressure (psi) X_2, Dwell time (s) X_3, Relative humidity (%) X_4, Vendor X_5, and Seal strength (psi) Y.

(a) Fit a multiple linear regression model to the data.
(b) Test for significance of regression using the analysis of variance, with $\alpha = 0.05$. What conclusions can you draw?

Case Study 2: (*Semiconductor study*)[2] An engineer at a semiconductor company wants to model the relationship between the device gain or *hFE* (Y) and the three parameters (independent variables): emitter-RS (X_1), base-RS (X_2), and emitter-to-base-RS (X_3). The data are shown below:

X_1 emitter-RS	X_2 base-RS	X_3 B-E-RS	Y *hFE*-1M-5 V
14.620	226.00	7.000	128.40
15.630	220.00	3.375	52.62
14.620	217.40	6.375	113.90
15.000	220.00	6.000	98.01
14.500	226.50	7.625	139.90

X_1 emitter-RS	X_2 base-RS	X_3 B-E-RS	Y *hFE*-1M-5 V
15.250	224.10	6.000	102.60
16.120	220.50	3.375	48.14
15.130	223.50	6.125	109.60
15.500	217.60	5.000	82.68
15.130	228.50	6.625	112.60
15.500	230.20	5.750	97.52
16.120	226.50	3.750	59.06
15.130	226.60	6.125	111.80
15.630	225.60	5.375	89.09
15.380	234.00	8.875	171.90
15.500	230.00	4.000	66.80
14.250	224.30	8.000	157.10
14.500	240.50	10.870	208.40
14.620	223.70	7.375	133.40

[2] Source: Myers and Montgomery (1995), used with permission.

(a) Fit a multiple linear regression model to the data.
(b) Predict the value of *hFE* to be taken at $X_{\ 1} = 14.5$, $X_{\ 2} = 220$, and $X_{\ 3} = 5.0$.
(c) Test for the significance of regression using analysis of variance with $\alpha = 0.05$. What conclusions can you draw?

Case Study 3: (*Electric consumption by a chemical plant*)[3] The electric power Y consumed each month by a chemical plant is thought to be related to the average ambient temperature X_1, the number of days in the month the plant was operating X_2, the average product purity X_3, and the tons of product produced X_4. The past year's historical data are available and are presented in the following table:

Y	X_1	X_2	X_3	X_4
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

(a) Fit a multiple linear regression model to the data.
(b) Predict the power consumption for a month in which $X_1 = 75°\text{F}$, $X_2 = 24$ days, $X_3 = 90\%$, and $X_4 = 98$ tons.

Case Study 4: (*Pipeline data collection*)[4] The Alaska pipeline data consists of in-field ultrasonic measurements of the depths of defects in the Alaska pipeline. The depth of the defects were then remeasured in the laboratory. These measurements were performed in six different batches. The data were analyzed to calibrate the bias of the field measurements relative to the laboratory measurements. These data were provided by Harry Berger, at the time a scientist in the Office of the Director of the Institute of Materials Research (now the Materials Science and Engineering Laboratory) of NIST. These data were used in a study conducted for the Materials Transportation Bureau of the US Department of Transportation. The variables observed are field defect size, lab defect size, and batch. The data for this case study are available on the book website: www.wiley.com/college/gupta/statistics2e.

Do the complete analysis of the data above. In the analysis, use the field measurement as the response variable, the laboratory measurement as the predictor variable, and batches as the qualitative predictor variable.

[3] Source: Myers and Montgomery (1995), used with permission.
[4] Source: Based on data from NIST and SEMATECH (2003)).

16.10 USING JMP

This section is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Consider the multiple linear regression model in four predictor variables, that is,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, 10$$

Use matrix notation to describe this model. Define the least-squares normal equations for this model. Assuming that X is a full-rank matrix, find the least-squares estimators for β.

2. In Problem 1, what are the dimensions of the HAT matrix $H = X(X'X)^{-1}X'$?

3. In Problem 2, let $H = X(X'X)^{-1}X'$ be the HAT matrix for the general regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Show that the HAT matrix H is idempotent. If I is an identity matrix, then show that $(I - H)$ is also an idempotent matrix.

4. In Problem 4 of Section 16.3, find a 95% prediction interval for Y when $X_1 = 75$, $X_2 = 70$, assuming that ε_i's are normally distributed with mean zero and variance σ^2.

5. Use the data in Problem 7 of Section 16.3 to
   (a) Fit a multiple linear regression model using X_1, X_3, X_4, and X_6 as the predictor variables;
   (b) Estimate the variance σ^2.

6. Find the inverse matrix $(X'X)^{-1}$ for the model in Problem 5.

7. Use the results of Problems 5(b) and 6, to find the variance and covariance matrix of the estimators $\hat{\beta}_3$ and $\hat{\beta}_4$ in the model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6$$

8. Refer to Problem 5 above. Assuming normality of the Y's,
   (a) Use a t-statistic to test each of the hypotheses at the 5% level of significance:

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0, \quad i = 1, 3, 4, 6$$

   (b) Find a 95% confidence interval for each of the β_i, $i = 1, 3, 4, 6$.

9. For the model in Problem 5, construct the ANOVA table and then use this ANOVA table to determine the value of R^2 and R^2_{adj}.

10. Repeat Problem 6 above for the model in Problem 7(a) of Section 16.3.

11. Refer to Problem 7 of Section 16.3. Assuming normality of the Y's,
    (a) Use the relevant t-statistic to test each of the hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0, i = 2, 3, 4, 5$. Use $\alpha = 0.05$.
    (b) Find a 95% confidence interval for each of the β_i, $i = 2, 3, 4, 5$.

12. Refer to Problem 1 of Practice Problems for Sections 16.6 and 16.7.
    (a) Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$ to the data of this problem.
    (b) Use the t-statistic to test each of the hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0, \quad i = 1, 2, 3, 4, 5$. Use $\alpha = 0.05$.
    (c) Find the p-value for each hypothesis in (b).

13. Refer to Problem 8 of Section 16.3. For the model in Problem 8, construct the ANOVA table and then use this ANOVA table to determine the value of R^2 and R^2_{adj}.

14. Refer to Problem 7 of Section 16.3.
    (a) Using MINITAB, find a best subset of predictor variables to fit the desired model.
    (b) Fit a model to these data using the stepwise technique and compare the results in (a) and (b) and see if your claim in (a) is valid.

15. Refer to Problem 7 of Section 16.3. Discuss whether or not one should use the model developed in Problem 7 to find a 95% confidence interval for $E(Y)$ and/or a 95% prediction interval for the pull strength Y at $X_1 = 8.0$, $X_3 = 38$, $X_4 = 80$, and $X_6 = 3.2$. Justify your answer.

16. Referring to Problem 11 of Section 16.3, find a 95% confidence interval for $E(Y)$ and a 95% prediction interval for Y when $X_1 = 0.25$ and $X_2 = 125$, assuming normality of the Y's.

17. A variable Y was observed at 12 different combinations of values of controlled variables X_1 and X_2, with the results shown below. If the relation between Y and (X_1, X_2) may be assumed to be linear in the region covered by the choice of the 12 pairs (X_1, X_2), fit the regression model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ to these data.

X_1	−2	−3	1	4	3	−2	−3	−1	−4	5	0	2
X_2	0.5	−0.5	−0.5	1.5	−2.5	−4.5	2.5	−3.5	0.5	1.5	1.5	3.5
Y	15	11	17	18	23	11	17	13	14	32	21	24

18. Refer to Problem 17. Set up the ANOVA table for the model in this problem. Use this ANOVA table to test the significance of the regression model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ at the 5% level of significance. Calculate the value of R^2 and R^2_{adj}.

19. An article in the *Journal of Pharmaceuticals Sciences* (vol. 80, 1991, 971–977) presents the following data on the observed mole fraction solubility of a solution at a constant

temperature, when the process is ran at certain values of dispersion, dipolar, and hydrogen bonding (Hansen partial solubility parameters):

Observations	Y	X_1	X_2	X_3	Observations	Y	X_1	X_2	X_3
1	0.22200	7.3	0.0	0.0	14	0.10100	7.3	2.5	6.8
2	0.39500	8.7	0.0	0.3	15	0.23200	8.5	2.0	6.6
3	0.42200	8.8	0.7	1.0	16	0.30600	9.5	2.5	5.0
4	0.43700	8.1	4.0	0.2	17	0.09230	7.4	2.8	7.8
5	0.42800	9.0	0.5	1.0	18	0.11600	7.8	2.8	7.7
6	0.46700	8.7	1.5	2.8	19	0.07640	7.7	3.0	8.0
7	0.44400	9.3	2.1	1.0	20	0.43900	10.3	1.7	4.2
8	0.37800	7.6	5.1	3.4	21	0.09440	7.8	3.3	8.5
9	0.49400	10.0	0.0	0.3	22	0.11700	7.1	3.9	6.6
10	0.45600	8.4	3.7	4.1	23	0.07260	7.7	4.3	9.5
11	0.45200	9.3	3.6	2.0	24	0.04120	7.4	6.0	10.9
12	0.11200	7.7	2.8	7.1	25	0.25100	7.3	2.0	5.2
13	0.43200	9.8	4.2	2.0	26	0.00002	7.6	7.8	20.7

Here Y is the negative logarithm of the mole fraction solubility, X_1 is the dispersion (Hansen partial solubility), X_2 is the dipolar partial solubility, and X_3 is the hydrogen bonding partial solubility.

(a) Fit the complete second-order regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \varepsilon$$

(b) Test for significance of regression, using $\alpha = 0.05$.

(c) Plot the residuals and comment on model adequacy.

(d) Test the hypothesis that the contribution of the second-order terms is zero, using $\alpha = 0.05$.

20. Consider the data in Problem 19.

(a) Use the stepwise regression method to fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

(b) Compute the Mallows' C_p, PRESS and R^2_{pred} statistics.

21. Refer to Problem 19. Fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \varepsilon$$

and compute the Mallows' C_p, PRESS and R^2_{pred} statistics. Compare these statistics with those computed in Problem 20 (b).

22. Refer to Problem 20. Calculate the hat matrix H and check that it is idempotent.

23. Use the hat matrix H obtained in Problem 22 to compute the residual error vector ε and the $Cov(\varepsilon)$ matrix. (Hint: See Equations (16.4.17) and (16.4.18.)

24. Refer to the data in Table 16.4.3 of Example 16.4.2.

(a) Fit the second-order regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \varepsilon$$

(b) Test for significance of regression model using the analysis of variance table with $\alpha = 0.05$. What are your conclusions?

25. Thirteen combinations of values of X_1, the amount of tricalcium aluminate (in percent) used in a mix of cement, and X_2, the amount of tricalcium (in percent) used in the mix, are used to study the effect on the heat involved, Y (in calories), during the hardening of the mix. The observations yield the following results (data from Hald, 1952):

$$\sum_{i=1}^{13} Y_i = 1240.5, \quad \sum_{i=1}^{13} X_{i1} = 97, \quad \sum_{i=1}^{13} X_{i2} = 626$$
$$\sum_{i=1}^{13} X_{i1}^2 = 1139, \quad \sum_{i=1}^{13} X_{i2}^2 = 33{,}050, \quad \sum_{i=1}^{13} X_{i1} X_{i2} = 4922$$
$$\sum_{i=1}^{13} X_{i1} Y_i = 10{,}032, \quad \sum_{i=1}^{13} X_{i2} Y_i = 62,027.8, \quad \sum_{i=1}^{13} Y_i^2 = 121{,}088$$

(a) Assuming the usual linear relationship, find the least-squares estimators of the coefficients of the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

(b) Determine 95% confidence intervals for $E(Y)$ and Y when $X_1 = 7, X_2 = 50$.

26. In Problem 25, estimate the variance covariance matrix of $\hat{\beta}$, the estimate of the parameter vector β.

27. Heat treating is often used to carbonize metal parts, such as gears. The thickness of the carburized layer is considered an important feature of the gear and contributes to the overall reliability of the part. Because of the critical nature of this feature, two different lab tests are performed on each furnace load. One test is run on a sample

pin that accompanies each load. The other test is a destructive test where an actual part is cross-sectioned. This test involves running a carbon analysis on the surface of both the gear pitch (top of the gear tooth) and the gear root (between the gear teeth). The following data are the results of the pitch carbon analysis test catch for 32 parts (Source: Data from Myers and Montgomery (1995); used with permission.).

X_1	X_2	X_3	X_4	X_5	Y	X_1	X_2	X_3	X_4	X_5	Y
1650	0.58	1.10	0.25	0.90	0.013	1650	2.20	1.10	1.10	0.80	0.024
1650	0.66	1.10	0.33	0.90	0.016	1650	2.20	1.10	1.10	0.80	0.025
1650	0.66	1.10	0.33	0.90	0.015	1650	2.20	1.15	1.10	0.80	0.024
1650	0.66	1.10	0.33	0.95	0.016	1650	2.20	1.10	1.10	0.90	0.025
1600	0.66	1.15	0.33	1.00	0.015	1650	2.20	1.10	1.10	0.90	0.027
1600	0.66	1.15	0.33	1.00	0.016	1650	2.20	1.10	1.50	0.90	0.026
1650	1.00	1.10	0.50	0.80	0.014	1650	3.00	1.15	1.50	0.80	0.029
1650	1.17	1.10	0.58	0.80	0.021	1650	3.00	1.10	1.50	0.70	0.030
1650	1.17	1.10	0.58	0.80	0.018	1650	3.00	1.10	1.50	0.75	0.028
1650	1.17	1.10	0.58	0.80	0.019	1650	3.00	1.15	1.66	0.85	0.032
1650	1.17	1.10	0.58	0.90	0.021	1650	3.33	1.10	1.50	0.80	0.033
1650	1.17	1.10	0.58	0.90	0.019	1700	4.00	1.10	1.50	0.70	0.039
1650	1.17	1.15	0.58	0.90	0.021	1650	4.00	1.10	1.50	0.70	0.040
1650	1.20	1.15	1.10	0.80	0.025	1650	4.00	1.15	1.50	0.85	0.035
1650	2.00	1.15	1.00	0.80	0.025	1700	12.50	1.00	1.50	0.70	0.056
1650	2.00	1.10	1.10	0.80	0.026	1700	18.50	1.00	1.50	0.70	0.068

Here the variables involved are Y: pitch; X_1: furnace temperature; X_2: duration of the carburizing cycle; X_3: carbon concentration; X_4: duration of the diffuse cycle; X_5: carbon concentration of the diffuse cycle.

(a) Fit the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

(b) Use an ANOVA table to test the significance of the regression model at the 5% level of significance.

28. Refer to data on heat treatment in Problem 27. Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

(a) Find 95% confidence intervals for each of the regression coefficients.

(b) Find a 95% confidence interval on mean pitch Y when

$$X_1 = 1675, \quad X_2 = 2.25, \quad X_3 = 1.12, \quad X_4 = 1.20, \quad X_5 = 0.85$$

(c) Find a 95% prediction interval on pitch Y when

$$X_1 = 1675, \quad X_2 = 2.25, \quad X_3 = 1.12, \quad X_4 = 1.20, \quad X_5 = 0.85$$

29. Refer to data on heat treatment in Problem 27.
    (a) Fit the linear regression model
    $$Y = \beta_0 + \beta_{12}X_1X_2 + \beta_{23}X_2X_3 + \varepsilon$$
    (b) Compute R^2_{pred} and comment on its value.

30. Refer to Problem 29.
    (a) Find 95% confidence intervals for each of the regression coefficients in the model in Problem 29(a).
    (b) Use a t-statistic to test the hypotheses $H_0 : \beta_{12} = 0$ versus $H_1 : \beta_{12} \neq 0$.

31. Refer to the data in Review Problem 19.
    (a) Fit to these data a complete second-order regression model using Stepwise regression method.
    (b) Estimate the variance of $\hat{Y}$ when $X_1 = 8.5, X_2 = 1$, and $X_3 = 6$.
    (c) Find the standard errors for each of the estimates of the regression coefficients in the model in part (a).
    (d) Find a 95% confidence interval for each of the regression coefficients.

32. Refer to Problem 29.
    (a) Find a 95% confidence interval for the mean pitch Y when $X_1 = 1670$, $X_2 = 2.20$, $X_3 = 1.15$.
    (b) Find a 95% prediction interval for pitch Y when $X_1 = 1670$, $X_2 = 2.20$, $X_3 = 1.15$.

33. Fifteen women are screened for breast cancer. The response variable is a binary variable Y ($Y = 1$ if a woman has breast cancer and $Y = 0$ if the woman does not have breast cancer) and the predictor variables X_1 (age), X_2 (age at first pregnancy), and X_3, a categorical predictor variable (1 if the woman has a family history of breast cancer and 0 if she does not have any family history). Results are given below. Using one of the statistical packages, fit a logistic regression model to these data and interpret your results:

Y	X_1	X_2	X_3
0	71	37	0
1	56	45	1
0	54	39	0
0	53	41	1
1	58	41	0
1	66	41	1
0	80	38	0
0	59	44	1
1	72	43	1
0	59	39	0
0	59	42	1
1	56	43	0
0	56	35	1
0	58	35	0
1	71	45	0

34. A study was performed to investigate the relationship between family income and ownership of a luxury car. Sixteen households were randomly selected and their family income X_1 (in units of 1000 dollars) and information about luxury car ownership ($Y = 1$ if family owns at least one luxury car; otherwise, $Y = 0$ were recorded). The data are shown below:

Y	X_1	Y	X_1
1	227	1	225
1	166	0	164
0	108	0	133
0	139	1	156
1	233	1	229
0	129	0	117
0	103	0	118
0	213	1	165

(a) Fit a logistic regression model to these data.
(b) Is the model you fitted in part (a) significant?
(c) Interpret the regression coefficient β_1 in terms of log odds ratio.
(d) Determine the estimated probability that a person with family income of $177,000 owns a luxury car.

35. An engineering society believes that the important factors that companies take into consideration in new hiring for senior positions are number of years of experience, X_1, and the number of publications/patents, X_2. The society selected 15 candidates who were interviewed recently and found to be or not to be hired. The data collected are given below, with Y the response variable ($Y = 1$ if the candidate was hired and 0 otherwise):

Y	1	1	0	1	0	0	1	1	1	0	0	1	1	1	0
X_1	12	15	9	9	10	7	16	15	8	6	10	16	14	20	12
X_2	39	38	26	20	30	22	10	37	38	16	22	36	28	34	35

(a) Fit a logistic regression model $\eta = \log(odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ to these data.
(b) Test each of the hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0, \quad i = 1, 2,$ at the 5% level of significance.

36. Refer to Problem 35.

(a) Determine the estimated probability that a person with 15 years' experience who has 30 papers/patents will be hired.
(b) Find a 95% confidence interval for each of the $\beta_i, i = 1, 2$.

37. A medical research team initiated a study to investigate the relationship between the stress level (Y) among university students and four other variables measured at the time of the study, namely X_1 age of the student, X_2 number of credit hours the student is taking, X_3 number of extracurricular activities the student is participating in, and a qualitative variable, X_4 gender. Note that gender is coded so that $X_4 = 0$ if the student is a male and 1 if the student is a female. The following data were collected on a simple random sample of 15 university students:

Y	133	55	95	38	128	124	68	106	131	41	80	94	116	67	41
X_1	22	25	21	19	28	22	20	23	21	30	31	27	23	21	22
X_2	16	12	12	15	16	16	14	14	16	12	14	14	16	12	12
X_3	3	2	0	2	4	3	1	3	2	0	1	2	2	1	1
X_4	1	1	0	0	1	0	0	1	1	1	0	0	1	0	0

(a) Fit a regression model $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

(b) Test, using the ANOVA table, the significance of regression at the 5% level of significance.

(c) Find a 95% confidence interval for mean stress level Y, and 95% prediction interval for stress level Y at $X_0 = (x_1, x_2, x_3, x_4) = (24, 14, 3, 1)$.

38. Refer to Problem 37. Consider the linear model using predictor variables X_1, X_2, and X_4. That is,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$$

(a) Fit the above regression model to the data in Problem 37.

(b) Construct an ANOVA table, and use it to test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_4 = 0$. Use $\alpha = 0.05$.

(c) Find a 95% confidence interval for each of the regression coefficients in the model for this problem.

39. Refer to Problem 38.

(a) Find a 95% confidence interval for $E(Y|X_0)$ at $X_0 = (24, 14, 1)$.

(b) Find a 95% prediction interval for Y at $X_0 = (24, 14, 1)$. Compare your result with the one you obtained in (a), and comment on these intervals.

40. Refer to data on stress level in Problem 37.

(a) Fit the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

(b) Compute R^2_{pred} for the models in Problems 37, 38, and (a) of this question and decide, based on the value of R^2_{pred}, which model fits the best.

41. Refer to Example 16.4.2. Re-analyze the data in Table 16.4.3 by using the stepwise regression technique. Use $\alpha = 0.15$, for the value of alpha (α) to enter a predictor variable. Compare the fitted model with the model in Example 16.4.2 and comment.

42. In Problem 41, it was found that the predictor variable X_3 is not important. Hence, reanalyze the data obtained by deleting the column corresponding to predictor variable X_3 from Table 16.4.3. Develop the ANOVA table for the new model and use an appropriate F-test to evaluate whether or not the fitted model is appropriate. Use $\alpha = 0.01$. Do you get the same model as in Problem 41?

Chapter 17

ANALYSIS OF VARIANCE

The focus of this chapter is the development of various experimental designs involving one or more factors having fixed effects, random effects, or mixed effects.

Topics Covered

- Design model and estimable functions
- One-way experimental layouts
- Multiple comparisons
- Determination of sample size
- Kruskal–Wallis test for one-way layouts (nonparametric method)
- Randomized complete block (RCB) design
- Friedman F_r-test for RCB designs
- Experiments with one missing observation in an experiment that used a RCB design
- Experiments with several missing observations in an experiment that used a RCB design
- Two-way experimental designs
- Two-way experimental layouts with one observation per cell
- Two-way experimental layouts with r (>1) observations per cell
- Blocking in two-way experimental designs
- Extending two-way experimental designs to n-way experimental designs
- Latin square designs
- Random effects model
- Mixed effects model
- Nested (hierarchical) design

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

Learning Outcomes

After studying this chapter, the reader will be able to

- Design and conduct various kinds of experiments in engineering, or other scientific fields, that involve one or more factors.
- Eliminate effects of one or more nuisance factors using appropriate experimental designs.
- Analyze data coming out of experiments with fixed, random, or mixed effects.
- Perform residual analysis to check the adequacy of the models under consideration.
- Use nonparametric techniques in certain kinds of designs when normality conditions are not valid.
- Summarize and interpret the results of these experiments.
- Estimate missing observations in certain kinds of designs, and to subsequently analyze the data as balanced data.
- Use statistical packages MINITAB, R, and JMP to analyze the data obtained by conducting these experiments.

17.1 INTRODUCTION

In many experiments, the main objective is to determine the effects of various factors on some response variable Y of basic or primary interest. For instance, in the study of abrasion resistance of a certain type of rubber, it may be important to determine the effect of chlorinating agents on such resistance. Or in the study of the strength of synthetic yarn, it may be important to determine the effects of viscosity of the molten form, rate of extrusion, and other factors on the strength Y of the yarn. In experiments such as these, it is important to design them carefully in terms of numbers of trials as well as choices of levels of the various factors involved.

Designed experiments, analyzed in accordance with certain principles discussed in this chapter, often make it possible to arrive at clearer and more trustworthy inferences about effects of the various factors. The principles of experimental design and methods of statistical analysis of experimental results considered in this chapter are commonly referred to as *analysis of variance* methods. This is a vast subject, and here we discuss only several of the commonly used experimental designs and their statistical analyses. The reader interested in pursuing this subject further should consult (Bennett and Franklin, 1954; Box et al., 1978; Cochran and Cox, 1957; Daniel, 1976; Dean and Voss, 1999; Hinkelmann and Kempthorne, 2005; Montgomery, 2009a,b), among the many other books devoted to the subject.

17.2 THE DESIGN MODELS

17.2.1 Estimable Parameters

Consider the following general linear model:

$$y_i = \mu + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{17.2.1}$$

As we saw in Chapter 16, the model (17.2.1) in matrix notation can be written as

$$Y = X\beta + \varepsilon \tag{17.2.2}$$

with

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where β is a vector of unknown parameters to be estimated and ε is a vector of random errors with all $\varepsilon_i's$ uncorrelated and having mean zero and common variance σ^2. The $[n \times (k+1)]$ matrix X contains, in the case of a regression model, preassigned values of the independent variables, and Y is a vector of observations on the dependent variable, sometimes called the vector of responses. In this section, we consider a special case of (17.2.1), where the matrix X, determined by the structure of the experimental design, consists of zeros and ones (see Example 17.2.1).

Example 17.2.1 (Fertilizers versus yield Y of a certain variety of wheat) *Suppose that we want to study the effect of two different fertilizers applied at a fixed level on the yield* Y *of a certain variety of wheat. This experiment is conducted by growing the given variety of wheat in a number of plots and applying fertilizer I in some plots selected randomly and fertilizer II in the remaining plots. (These plots are assumed to be homogeneous with respect to all other characteristic variations and all other known factors.) The response in this experiment is the yield of wheat from each plot.*

For the experiment above, we consider the following design model:

$$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n_j, j = 1,\ 2 \tag{17.2.3}$$

where Y_{ij} is the observed yield of wheat in the ith plot when the jth fertilizer is used. Here, β_j is the effect due to using the jth fertilizer and ε_{ij} is the random error due to all uncontrolled and unknown factors. Now, if we suppose that each of the two fertilizers is applied only to two plots, then the model (17.2.3) can be more explicitly written as

$$y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1, 2;\ j = 1, 2 \tag{17.2.4}$$

which, in matrix notation, can be rewritten as

$$Y = X\beta + \varepsilon \tag{17.2.5}$$

where

$$Y = \begin{bmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \end{bmatrix}$$

Here, the vector Y is the vector of observations or responses (yields) when the two fertilizers are used, β is the vector of unknown parameters, ε is the error vector, and X is the *design matrix*. We may note here that the design matrix X is a 4 × 3 matrix with 0s and 1s and is of rank 2, since the sum of the last two columns is equal to the first column. We are interested in estimating β by using one of the standard techniques, namely the method of least squares. As discussed in Chapter 16, this method calls for the minimization of the error sum of squares $\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$. This leads us to the following least-squares normal equations:

$$X'X\hat{\beta} = X'Y \tag{17.2.6}$$

where $\hat{\beta}$ denotes the least-squares estimator of β.

As mentioned earlier, X is not of full rank, which means that the (3×3) matrix $X'X$ is not of full rank, so that $(X'X)^{-1}$ does not exist, and hence, the solution of (17.2.6) is not unique. However, by using a generalized inverse G of $X'X$, we can find a solution for β. We do not include any discussion of the generalized inverse of a matrix because it is beyond the scope of this book, but proceed as follows.

When the design matrix is not of full rank, it is not possible to determine unique estimates of all parameters separately unless we impose some *side conditions* on the model (to be studied later). However, we can estimate certain linear combinations of parameters. For instance, in the example above, we can find unique estimates of $E(Y_{i1}) = \mu + \beta_1$ and $E(Y_{i2}) = \mu + \beta_2$. This is achieved by using the technique of *reparameterization*, which transforms a design matrix of less than full rank to the design matrix of full rank. Then we can use the theory of general linear models for the full rank case, as we did in Chapter 16. For example, in (17.2.4) we consider $\mu + \beta_j = \mu_j, j = 1, 2$, so we can rewrite (17.2.4) as

$$y_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1, 2, j = 1, 2 \tag{17.2.7}$$

In model (17.2.7), the vector γ of parameters and the design matrix X are

$$\gamma = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The design matrix X is now of full rank, which means that we can find a unique solution of the system $X'X\hat{\gamma} = X'Y$. In other words, we can determine the unique estimates of μ_1 and μ_2.
Suppose that, the observation vector is

$$Y = \begin{bmatrix} 14 \\ 10 \\ 12 \\ 16 \end{bmatrix}$$

Then (the reader should verify) we can easily see that $\hat{\mu}_1 = 12$, and $\hat{\mu}_2 = 14$.

17.2.2 Estimable Functions

The discussion on estimable functions is not included in this book, but is available for download from the book website: www.wiley.com/college/gupta/statistics2e. After consulting this website, the reader should attempt the following practice problems.

PRACTICE PROBLEMS FOR SECTION 17.2

1. Consider the design model $Y = X\beta + \varepsilon$, where X is of the order $n \times m$ $(m \leq n)$ and rank (X) $= r < m$. Show that $X\beta$ is a set of estimable functions.
2. Refer to Problem 1. Show that the set $X\beta$ contains r linearly independent estimable functions.
3. Suppose that an agronomist wants to study the effect of two fertilizers (each applied using a fixed amount) on the yield of a certain variety of corn. He makes two observations using each fertilizer, and writes the model for this experiment as
$$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad i = 1,\ 2,\ j = 1,\ 2$$
Show that $c_1\beta_1 + c_2\beta_2$, where $c_1 + c_2 = 0$, is a linearly estimable function.
4. Refer to Problem 3 above. Show that $\beta_1 + \beta_2$ is not an estimable function.
5. Refer to Problem 3 above. Is it possible to find in this problem four linearly independent estimable functions?
6. Consider the model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, i = 1, 2, 3, j = 1, 2,$ where $\varepsilon \sim N(0, \sigma^2 I)$, $\varepsilon = (\varepsilon_{11}\ \varepsilon_{21}\ \varepsilon_{31}\ \varepsilon_{12}\ \varepsilon_{22}\ \varepsilon_{32}\ )'$. Suppose the model above is written in matrix notation, $Y = X\beta + \varepsilon$.
   (a) Find the rank of the matrix X.
   (b) Find two linearly independent estimable functions.

17.3 ONE-WAY EXPERIMENTAL LAYOUTS

17.3.1 The Model and Its Analysis

In Sections 9.6 and 9.7, we discussed the problem of testing hypotheses about the difference of two normal populations on the basis of samples from both these populations. We saw that if the two population variances are unknown but could be assumed equal, the statistic used for making the test is a Student t-statistic that involves the two sample averages and the two sample variances.

Now suppose that we have an experiment involving samples from three or more populations. The question of how to test a hypothesis concerning the means of these populations from which the samples are assumed to have been drawn is considered in this section.

As we noted in Section 17.2.1, each experiment gives rise to some technical terms such as response, factors, and levels. For example, suppose that we are interested in comparing the yields per plot of different varieties of corn. Then, the yield per plot is the *response*, the variety of corn is the *factor*, and different varieties of corn are the *levels* of this factor. Here, plots are the *experimental units.*

Now suppose that $A_1, A_2, \ldots, A_a$ are different levels of a factor A, and we want to study the effects of $A_1, A_2, \ldots, A_a$ on some response variable Y of primary interest. We set up an experiment in which n_1 observations are made on Y when level A_1 is present, n_2 observations are made on Y when level A_2 is present, and so on. The levels of A are often referred to as the "treatments," there being a treatments in this experimental design. Such designs are also called completely randomized designs when the treatments are allocated to the experimental units in a completely random fashion. Further, the experimental units are assumed to be homogeneous with respect to all known factors.

Alternatively, we could consider the following simple model for such an experiment. Suppose that $(y_{11}, y_{21}, \ldots, y_{n_1 1})$ is a random sample of size n_1 from a population having the normal distribution $N(\mu + \delta_1, \sigma^2)$, $(y_{12}, y_{22}, \ldots, y_{n_2 2})$ is a random sample of size n_2 from a population having the normal distribution $N(\mu + \delta_2, \sigma^2), \ldots,$ $(y_{1a}, y_{2a}, \ldots, y_{n_a a})$ is a random sample of size n_a from a population having the normal distribution $N(\mu + \delta_a, \sigma^2)$. Here, we carry out the experiment in a manner that the samples are all independent and $(\mu,\ \delta_1, \ldots,\ \delta_a,\ \sigma^2)$ are unknown parameters with $\delta_1, \ldots, \delta_a$ satisfying the condition

$$n_1\delta_1 + n_2\delta_2 + \cdots + n_a\delta_a = 0 \tag{17.3.1}$$

The parameters $\delta_1, \delta_2, \ldots, \delta_a$ are referred to as the effects of treatments $A_1, A_2, \ldots, A_a$, respectively.

The parameter μ is sometimes called the *overall (population) mean*, and σ^2 is the common variance of $Y_{ij}, i = 1, \ldots, n_j, j = 1, \ldots, a$. The samples may be arranged as shown in Table 17.3.1, and we now write the model as

$$Y_{ij} = \mu + \delta_j + \varepsilon_{ij}, \quad i = 1, \ldots, n_j, j = 1, \ldots, a \tag{17.3.2}$$

where

$$\sum_{j=1}^{a} n_j\delta_j = 0 \tag{17.3.3}$$

Table 17.3.1 One-way experiment layout factor A-levels.

Levels	A_1	A_2	$\cdots$	A_j	$\cdots$	A_a	
	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1a}	
	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2a}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$y_{n_1 1}$	$y_{n_2 2}$	$\cdots$	$y_{n_j j}$	$\cdots$	$y_{n_a a}$	
Total	$\sum_{i=1}^{n_1} y_{i1} = T_1$	$\sum_{i=1}^{n_2} y_{i2} = T_2$	$\cdots$	$\sum_{i=1}^{n_j} y_{ij} = T_j$	$\cdots$	$\sum_{i=1}^{n_a} y_{ia} = T_a$	$G = \sum_{j=1}^{a} T_j$

The *restriction* or *side-condition* on the δ_j in (17.3.3) enters quite naturally and ensures that the parameters $(\mu,\ \delta_1, \ldots,\ \delta_a)$ are uniquely defined (see Section 17.2). In the model (17.3.2), it is not necessary that the sample sizes be equal, but in practice, we always prefer that all the sample sizes be equal. If all sample sizes are equal, then data are called balanced; otherwise, the data are called unbalanced. The data obtained from a one-way experiment are recorded as shown in Table 17.3.1.

In model (17.3.2), Y_{ij} are observable random variables and ε_{ij} are unobservable random variables that are assumed to be independently and identically distributed as $N(0, \sigma^2)$. Hence, Y_{ij} is normally distributed with mean $\mu + \delta_j$ and variance σ^2. Indeed (17.3.2) implies that Y_{ij}, apart from the normally distributed random error ε_{ij}, is made up of an overall constant (mean) μ plus the effect δ_j of the treatment Aj used to generate $Y_{ij}, i = 1, \ldots, n_j$

Now let $\overline{y}_{.j}$ be the average of the *j*th sample and $\overline{y}_{..}$ be the average of all the samples pooled together. That is,

$$\overline{y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} = \frac{T_j}{n_j} \tag{17.3.4}$$

$$\bar{y}_{..} = N^{-1}\sum\sum y_{ij} = \frac{G}{N}, \quad G = \sum\sum y_{ij} \tag{17.3.5}$$

where $\sum\sum$ stands for $\sum_{j=1}^{a}\sum_{i=1}^{n_j}$, with the total sample size N given by

$$N = \sum_{j=1}^{a} n_j \tag{17.3.6}$$

The reader can easily verify that the least square estimates of the overall mean μ and the treatment effects δ_j $(j = 1, 2, \ldots, a)$ are given by

$$\hat{\mu} = \bar{y}_{..} \tag{17.3.7}$$

$$\hat{\delta}_j = \bar{y}_{.j} - \bar{y}_{..} \tag{17.3.8}$$

Under the normality assumption, these estimates are also maximum likelihood estimates. With or without the normality assumption, from (17.3.2) and (17.3.3), it can be seen that $E(\hat{\delta}_j) = \delta_j$.

Further, it can be seen that the total variation SS_{total} of all observations in the pooled sample can be written in two ways. First,

$$SS_{total} = \sum\sum(y_{ij} - \bar{y}_{..})^2 = \sum\sum y_{ij}^2 - N\bar{y}_{..}^2 = \sum\sum y_{ij}^2 - \frac{G^2}{N} \tag{17.3.9}$$

Indeed, SS_{total} is often called the corrected sum of squares, with the implication that we have subtracted the *correction term* or *correction factor* $N\bar{y}_{..}^2$ from the sum of squares of all the observations.

However, we also can write SS_{total} as follows:

$$\begin{aligned} SS_{total} &= \sum\sum(y_{ij} - \bar{y}_{..})^2 = \sum\sum[(y_{ij} - \bar{y}_{.j}) + (\bar{y}_{.j} - \bar{y}_{..})]^2 \\ &= \sum\sum(y_{ij} - \bar{y}_{.j})^2 + \sum\sum(\bar{y}_{.j} - \bar{y}_{..})^2 \end{aligned} \tag{17.3.10}$$

since the sum of the cross products vanishes, that is,

$$\sum\sum(y_{ij} - \bar{y}_{.j})(\bar{y}_{.j} - \bar{y}_{..}) = 0 \tag{17.3.11}$$

In the second line of (17.3.10), we define

$$SS_E = \sum\sum(y_{ij} - \bar{y}_{.j})^2 \tag{17.3.12}$$

$$SS_A = \sum\sum(\bar{y}_{.j} - \bar{y}_{..})^2 = \sum_{j=1}^{a} n_j(\bar{y}_{.j} - \bar{y}_{..})^2 \tag{17.3.13}$$

Thus we have

$$SS_{total} = SS_E + SS_A \tag{17.3.14}$$

which means that we have broken down the total variation SS_{total} into two parts: SS_A, which reflects the variation between samples, or the variation due to the various levels of A, and is usually referred to as the *between samples sum of squares* (or simply *between sum of squares* or *treatment sum of squares*), and SS_E, which reflects variation within samples and is usually called the *within-samples sum of squares* (or *simply within sum of squares* or *error sum of squares*).

To look at the breakdown differently, we remind ourselves that we consider the Y_{ij} to be generated according to the model (17.3.2). This means that

$$\bar{y}_{\cdot j} = \mu + \delta_j + \bar{\varepsilon}_{\cdot j} \tag{17.3.15}$$

$$\bar{y}_{\cdot\cdot} = \mu + \bar{\varepsilon}_{\cdot\cdot} \tag{17.3.16}$$

Using (17.3.15) and (17.3.16), we can write SS_A and SS_E as

$$SS_E = \sum\sum(\varepsilon_{ij} - \bar{\varepsilon}_{\cdot j})^2 \tag{17.3.17}$$

$$SS_A = \sum_{j=1}^{a} n_j(\delta_j + \bar{\varepsilon}_{\cdot j} - \bar{\varepsilon}_{\cdot\cdot})^2 \tag{17.3.18}$$

Note that SS_E depends only on the random variables ε_{ij}, which are $N(0, \sigma^2)$ variables, and SS_A depends on the parameters $\delta_1, \ldots, \delta_a$ as well as on the ε_{ij}. The reader may in fact verify that

$$E(SS_E) = \sigma^2 \sum_{j=1}^{a}(n_j - 1) = \sigma^2(N - a) \tag{17.3.19}$$

and

$$E(SS_A) = \sigma^2(a - 1) + \sum_{j=1}^{a} n_j \delta_j^2 \tag{17.3.20}$$

Note that the expectation of SS_E does not depend on $\delta_1, \ldots, \delta_a$.

Now, if $\delta_j = 0, j = 1, \ldots, a$, then, since the ε_{ij}s are normally distributed, the two quantities

$$\frac{SS_E}{\sigma^2}, \quad \frac{SS_A}{\sigma^2} \tag{17.3.21}$$

are independent random variables having chi-square distribution with $N - a$ and $a - 1$ degrees of freedom, respectively. It follows from Definition 17.3.4 that the ratio

$$\frac{SS_A/(a-1)\sigma^2}{SS_E/(N-a)\sigma^2} = \frac{SS_A/(a-1)}{SS_E/(N-a)} = \frac{MS_A}{MS_E} \tag{17.3.22}$$

is distributed as Snedecor's F-distribution (noncentral) with $(a - 1,\ N - a)$ degrees of freedom, and noncentral parameter $\sum_{j=1}^{a} n_j \delta_j^2/(a-1)$.

To test the null hypothesis of zero effects due to treatments $A_1, \ldots, A_a$ of the A factor, that is,

$$H_0 :\ \delta_1 = \cdots = \delta_a = 0 \text{ versus } H_1 :\ \delta_1, \ldots, \delta_a \text{ are not all } 0 \tag{17.3.23}$$

then this test may be performed by using the ratio in (17.3.22), which under H_0 is distributed as the central $F_{a-1,N-a}$ random variable. More precisely, we reject H_0 at the α level of significance if the observed value of (17.3.22) is such that

$$\frac{MS_A}{MS_E} > F_{a-1,N-a;\ \alpha} \tag{17.3.24}$$

Otherwise, we do not reject H_0.

If we reject H_0, which means that we reject the hypothesis that the effects $\delta_1, \ldots, \delta_a$ due to $A_1, \ldots, A_a$ are all zero, we estimate the effects $\delta_1, \ldots, \delta_a$ from (17.3.7) and (17.3.8) we have

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\delta}_j = \bar{y}_{.j} - \bar{y}_{..}, \quad i = 1, 2, \ldots, a, \tag{17.3.25}$$

and $\hat{\mu} = \bar{y}_{..}$ and $\hat{\delta}_j = \bar{y}_{.j} - \bar{y}_{..}$ are unbiased estimators for μ and δ_j, respectively. Furthermore, note that

$$n_1\hat{\delta}_1 + \cdots + n_a\hat{\delta}_a = 0$$

Of course, whether or not H_0 of (17.3.23) is true, from (17.3.19) we see that $S^2 = SS_E/(N-a) = MS_E$ is an unbiased estimator for σ^2. We could collect all of the constituents of our analysis shown so far into what is called an *analysis of variance table*, or simply the ANOVA table, as given below in Table 17.3.2. Note that under H_0, $E(MS_A) = \sigma^2$. This means that if H_0 is true, we can expect a "low value" of MS_A/MS_E, that is, a value around 1, which further justifies the procedure outlined in (17.3.24).

Table 17.3.2 ANOVA table for a one-way experimental layout.

Source of variation	Sum of squares	Degrees of freedom	Mean square	Expected mean square	F-ratio test
Between samples	SS_A	$a-1$	$MS_A = \dfrac{SS_A}{a-1}$	$\sigma^2 + \dfrac{1}{a-1}\sum_{j=1}^{a} n_j\delta_j^2$	$\dfrac{MS_A}{MS_E}$
Within samples	SS_E	$N-a$	$MS_E = \dfrac{SS_E}{N-a}$	σ^2	
Total	SS_{total}	$N-1$			

Estimators for $\delta_1, \ldots, \delta_a$ if H_0 is rejected are as given by (17.3.25), that is,

$$\hat{\delta}_j = \bar{y}_{.j} - \bar{y}_{..}, \quad j = 1, 2, \ldots, a$$

Estimator for σ^2 is

$$S^2 = \frac{SS_E}{N-a} = MS_E$$

Estimator for μ is

$$\hat{\mu} = \bar{y}_{..}$$

A convenient computational scheme for finding the relevant sums of squares in Table 17.3.2 proceeds as follows: Denote by T_j and SS_j the total and sum of squares, respectively, of the observations on the jth treatment, that is,

$$T_j = \sum_{i=1}^{n_j} y_{ij} \quad \text{and} \quad SS_j = \sum_{i=1}^{n_j} y_{ij}^2 \tag{17.3.26}$$

Using the notation of Table (17.3.1), denote the grand total by

$$G = \sum_{j=1}^{a} T_j$$

Then write the sum of squares entries in Table 17.3.2 as

$$SS_A = \sum_{j=1}^{a} \frac{T_j^2}{n_j} - \frac{G^2}{N}, \quad SS_E = SS_{total} - SS_A, \quad SS_{total} = \sum_{j=1}^{a} SS_j - \frac{G^2}{N} \qquad (17.3.27)$$

where T_j and SS_j are defined in (17.3.26).

Example 17.3.1 (Thermometers) *Four thermometers labeled 1, 2, 3, and 4 were used to make determinations Y of the melting point of hydroquinine in degrees centigrade, with the results as shown in Table 17.3.3. The experiment was carried out in random order.*

We want to test at the 5% level of significance the hypothesis that there is no significant variation in the means of the melting points as determined by the four thermometers. (The factor in this experiment is "thermometer," the levels are the four different thermometers, and their effects are $\delta_1, \delta_2, \delta_3,$, and δ_4, respectively.)

Note that if we code the observations by merely subtracting a constant from each observation, then the sums of squares of deviations are not affected. Hence, we analyze values of the variable $Y - 170$. This gives the values of $y - 170$ and $(y - 170)^2$, respectively, as shown in Tables 17.3.4 and 17.3.5.
Now, by using (17.3.27), we find that

$$\begin{aligned} SS_{total} &= (4.0)^2 + (3.0)^2 + \cdots + (1.0)^2 - \frac{(28.5)^2}{11} \\ &= 84.75 - 73.8409 = 10.9091 \\ SS_A &= \textstyle\sum_{j=1}^{4} \frac{T_j^2}{n_j} - \frac{G^2}{N} \\ &= \frac{(13.5)^2}{4} + \frac{(5)^2}{2} + \frac{(5.5)^2}{3} + \frac{(4.5)^2}{2} - \frac{(28.5)^2}{11} \\ &= 78.2708 - 73.8409 = 4.4299 \end{aligned}$$

and

$$SS_E = SS_{total} - SS_A = 10.9091 - 4.4299 = 6.4792$$

The analysis of variance table for the data of Table 17.3.3, appears in Table 17.3.6. The upper 5% point of $F_{3,7}$ is $F_{3,7;0.05} = 4.347$. The observed value of the F-statistic is 1.595, which is smaller than 4.347, so we do not reject the null hypothesis and conclude that there are no significant differences (at the 5% level) between the thermometers.

Table 17.3.3 Melting point of hydroquinine.

1	2	3	4
174.0	173.0	171.5	173.5
173.0	172.0	171.0	171.0
173.5		173.0	
173.0			

Table 17.3.4 Coded data for Example 17.3.1 ($y_{ij} - 170$).

	1	2	3	4	
	4.0 3.0 3.5 3.0	3.0 2.0	1.5 1.0 3.0	3.5 1.0	
Total	13.5	5.0	5.5	4.5	28.5

Table 17.3.5 $(y_{ij} - 170)^2$ squared coded observations.

	1	2	3	4	
	16.00 9.00 12.25 9.00	9.00 4.00	2.25 1.00 9.00	12.25 1.00	
Total	46.25	13.00	12.25	13.25	84.75

Table 17.3.6 ANOVA table for the data in Table 17.3.3.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-test
Between thermometers	4.4299	$a - 1 = 3$	1.4766	$\frac{1.4766}{0.9256} = 1.595$
Within samples	6.4792	$N - a = 7$	0.9256	
Total	$SS_{total} = 10.9091$	$N - 1 = 10$		

17.3.2 Confidence Intervals for Treatment Means

From the model (17.3.2), we see that the expectation of the jth treatment mean is given by

$$\mu_j = \mu + \delta_j, \quad j = 1, 2, \ldots, a \tag{17.3.28}$$

From Equations (17.3.7) and (17.3.8), it follows that a point estimator of μ_j is given by

$$\hat{\mu}_j = \hat{\mu} + \hat{\delta}_j = \bar{y}_{..} + \bar{y}_{.j} - \bar{y}_{..} = \bar{y}_{.j} \tag{17.3.29}$$

Assuming that ε_{ij} are normally distributed, it can be easily shown that the $\bar{y}_{.j}$'s are independently and normally distributed as $N(\mu_j,\ \sigma^2/n_j)$. For balanced data ($n_1 = \cdots = n_a = n$), this distribution becomes $N(\mu_j,\ \sigma^2/n)$. Since σ^2 is usually unknown, we use MS_E as an estimator of σ^2 and the Student t-distribution to show that a 100 $(1 - \alpha)$%

confidence interval for μ_j is given by

$$\left(\overline{y}_{\cdot j} \pm t_{N-a;\ \alpha/2}\sqrt{\frac{MS_E}{n_j}}\right) \tag{17.3.30}$$

For the balanced data, a 100 $(1-\alpha)$% confidence interval is given by $(N = na, n_j = n)$

$$\left(\overline{y}_{\cdot j} \pm t_{N-a;\ \alpha/2}\sqrt{\frac{MS_E}{n}}\right) \tag{17.3.31}$$

A $100(1-\alpha)$% confidence interval for the difference of two treatment means $\mu_i - \mu_j$ is given by

$$\left((\overline{y}_{\cdot i} - \overline{y}_{\cdot j}) \pm t_{N-a;\ \alpha/2}\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\right) \tag{17.3.32}$$

and for the balanced data, a 100 $(1-\alpha)$% confidence interval is given by

$$\left((\overline{y}_{\cdot i} - \overline{y}_{\cdot j}) \pm t_{N-a;\ \alpha/2}\sqrt{MS_E\left(\frac{2}{n}\right)}\right) \tag{17.3.33}$$

Example 17.3.2 (Example 17.3.1, revisited) *Refer to the thermometers data in Example 17.3.1. Find point estimates for the treatment effects and find a 95% confidence interval for mean of treatment 1. Find a 95% confidence interval for the difference of treatment 1 and treatment 4.*

Solution: We first remind the reader that the coded data in Table 17.3.4 was found by subtracting 170 from each of the observations in Table 17.3.3. Hence to find, for example, $\overline{y}_{\cdot 1}$, we use Table 17.3.4 with the coding to arrive at $\overline{y}_{\cdot 1} = 13.5/4 + 170 = 3.375 + 170 = 173.375$, and so on.

Now from Equation (17.3.25), we obtain

$$\begin{aligned}
\hat{\delta}_1 &= \overline{y}_{\cdot 1} - \overline{y}_{\cdot\cdot} = 173.375 - 172.59 = 0.785\\
\hat{\delta}_2 &= \overline{y}_{\cdot 2} - \overline{y}_{\cdot\cdot} = 172.50 - 172.59 = -0.09\\
\hat{\delta}_3 &= \overline{y}_{\cdot 3} - \overline{y}_{\cdot\cdot} = 171.833 - 172.59 = -0.757\\
\hat{\delta}_4 &= \overline{y}_{\cdot 4} - \overline{y}_{\cdot\cdot} = 172.25 - 172.59 = -0.34
\end{aligned}$$

Using (17.3.30), a 95% confidence interval for μ_1, the mean of treatment 1 based on n_1 observations, is given by $(t_{7;\,0.025} = 2.365)$

$$173.375 \pm 2.365\sqrt{0.9256/4} = 173.375 \pm 1.138$$

Thus, a 95% confidence interval for the mean of treatment 1 is

$$(172.237, 174.513)$$

Now, using the result in (17.3.32), we obtain a 95% confidence interval for the difference of treatment 1 and treatment 4 as

$$(173.375 - 172.25) \pm 2.365\sqrt{0.9256\left(\frac{1}{4} + \frac{1}{2}\right)} = 1.125 \pm 1.97 = (-0.845, 3.095)$$

We can also analyze data of a one-way layout experiment using MINITAB, R, or JMP.

Example 17.3.3 (Onion rings) *Five different types of oil (olive, soybean, corn, peanut, and sunflower) are often used for frying onion rings. It is not known whether the amount of oil absorbed by the onion rings depends on the type of oil. For five types of oil, certain batches of equal size (6) of onion rings are prepared. The experiment was carried out in random order. The data in Table 17.3.7 show the amount of oil (in grams) absorbed per batch. We want to test a hypothesis at the 5% level of significance that the absorption of oil in frying onion rings is same for all five types of oil.*

MINITAB

The experimental design model we use for this experiment is

$$Y_{ij} = \mu + \delta_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, 6, \; j = 1, 2, \ldots, 5$$

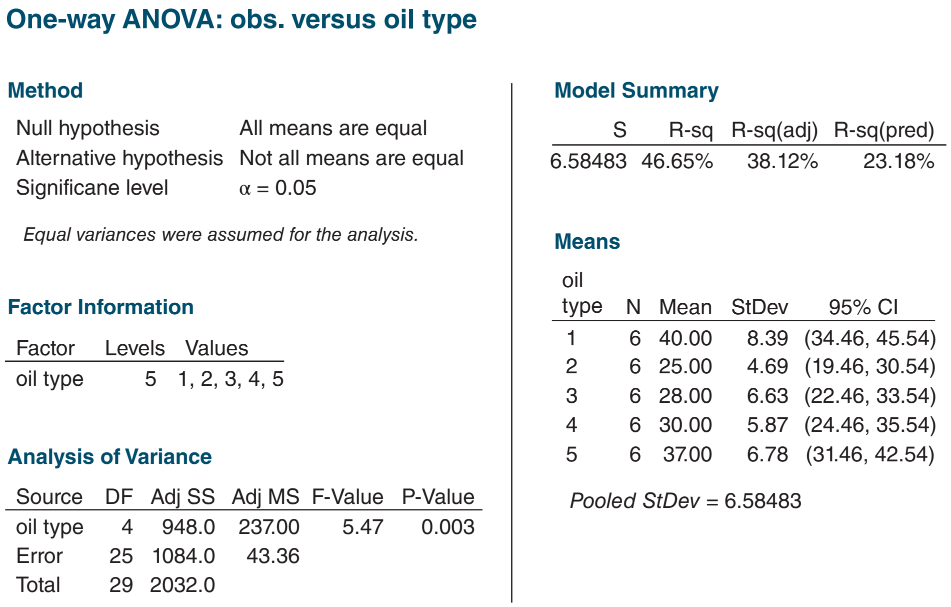
To analyze the data in Table 17.3.7, using MINITAB, we proceed as follows:

1. Enter the data for all five samples or treatments (oils) in column C_1.
2. In column C_2, enter the sample identifiers, say 1 for olive oil, 2 for soybean, and so on. We name the two columns obs. and oil type, respectively.
3. From bar menu select **Stat** > **ANOVA** > **One-way....**
4. In the dialog box that appears, select **Response data are in one column for all factors** and type in obs. in the box next to **Response** and oil type in the box next to **Factor**.

Table 17.3.7 Coded data ($y - 100$) in grams of oil absorbed per batch.

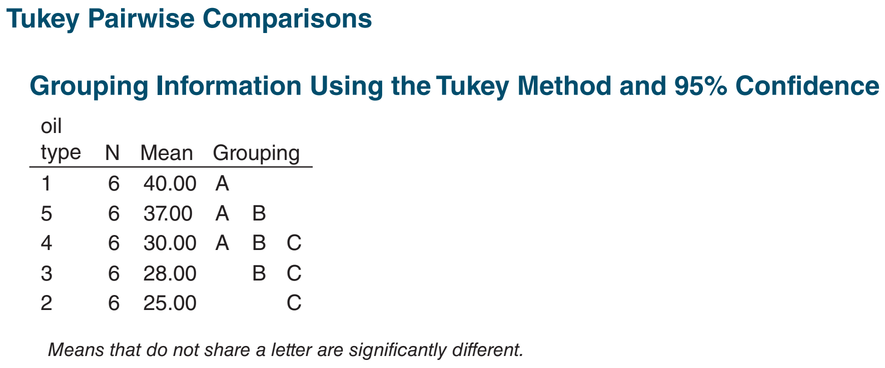
Oil Batch	Olive	Soybean	Corn	Peanut	Sunflower
1	47	22	29	23	42
2	33	19	24	37	31
3	41	25	25	24	47
4	30	31	20	29	39
5	52	23	31	36	30
6	37	30	39	31	33
Totals	240	150	168	180	222

5. Click **Options**, then for **Confidence level**, select 95.0, since we want to test the hypothesis at the 5% level of significance.
6. Use the graph options to select Four in one and click **OK**. Again click **OK**.
7. The output appears in the Session window as shown here.

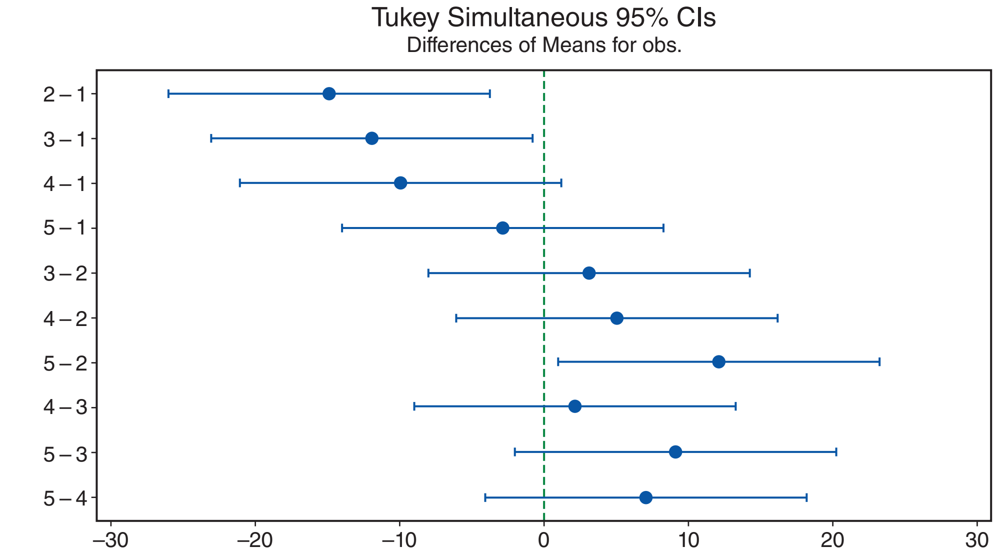


```
One-way ANOVA: obs. versus oil type

Method

 Null hypothesis         All means are equal
 Alternative hypothesis  Not all means are equal
 Significane level       α = 0.05

  Equal variances were assumed for the analysis.

Factor Information

 Factor    Levels  Values
 oil type       5  1, 2, 3, 4, 5

Analysis of Variance

 Source    DF  Adj SS  Adj MS  F-Value  P-Value
 oil type   4   948.0  237.00     5.47    0.003
 Error     25  1084.0   43.36
 Total     29  2032.0

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
 6.58483  46.65%     38.12%      23.18%

Means

 oil
 type  N   Mean  StDev  95% CI
 1     6  40.00   8.39  (34.46, 45.54)
 2     6  25.00   4.69  (19.46, 30.54)
 3     6  28.00   6.63  (22.46, 33.54)
 4     6  30.00   5.87  (24.46, 35.54)
 5     6  37.00   6.78  (31.46, 42.54)

  Pooled StDev = 6.58483
```

Tukey 95% simultaneous confidence intervals, for example can be obtained by selecting **Stat** > **ANOVA** > **One-way . . .** > **Comparisons . . .** > **Error rate for comparison** > **Tukey**, family error rate and entering the value of significance level (see Section 17.3.3). All pairwise comparisons among levels of oil type are



```
Tukey Pairwise Comparisons

 Grouping Information Using the Tukey Method and 95% Confidence

  oil
  type  N   Mean  Grouping
  1     6  40.00  A
  5     6  37.00  A  B
  4     6  30.00  A  B  C
  3     6  28.00     B  C
  2     6  25.00        C

   Means that do not share a letter are significantly different.
```



If an interval does not contain zero, the corresponding means are significantly different.

Since the p-value in the ANOVA table is 0.003, smaller than 0.05, we reject the null hypothesis that the absorption of oil in frying onion rings is the same for all five types of oil. And since we have rejected the hypothesis that the absorption of oil in frying onion rings is the same for all five oil types, we would like to find the estimates and confidence intervals for the treatment means $\mu_1, \ldots, \mu_5$, where $\mu_j = \mu + \delta_j \quad (i = 1, \ldots, 5)$. These estimates and confidence intervals are given after the ANOVA table. Note that the pooled standard deviation is $S = 6.585$. The Tukey's simultaneous confidence intervals for all pairs given in the printout above will be studied in more depth in the next section.

USING R

Solution: The R function ‘aov()’ can be used to fit the required ANOVA model as shown in the following R-code. Also, the R function ‘TukeyHSD()’ can be used to conduct Tukey's multiple comparison tests.

```
obs = c(47,33,41,30,52,37,22,19,25,31,23,30,29,24,25,20,31,39,23,37,24,29,36,31,42,
31,47,39,30,33)
oil.type = c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,5,5,5,5,5,5)

#Fitting ANOVA model. Make sure to change variable ‘oil.type’ to a ‘Factor’ as
follows:
model = aov(obs ~ factor(oil.type))

#ANOVA output
anova(model)
```

```
#Diagnostic plots
par(mfrow=c(2,2))
plot(model)

#Tukey's test
TukeyHSD(model)
```

The R output (not shown here) is identical to that obtained from the MINITAB procedure. Based on this output, and as concluded earlier from the MINITAB output, we have ample evidence to reject the hypothesis that all five oil means are equal.

We now turn to verifying the Model Adequacy:

1. The graph of residuals $[(y_{ij} - \overline{y}_{.j}), i = 1, \ldots, n_j; j = 1, \ldots, a]$ versus order shows that the assumption that residuals are independent is reasonable (see Figure 17.3.1).
2. In the normal probability plot in Figure 17.3.1, almost all the residuals fall on a straight-line fairly well, which implies that the assumption of normality is valid. (Here fits are $\hat{\mu} + \hat{\delta}_j = \overline{y}_{..} + \overline{y}_{.j} - \overline{y}_{..} = \overline{y}_{.j},\ j = 1, \ldots, 5.$)
3. The graph of residuals versus fits illustrates the assumption that equal variances show no abnormalities.

We conclude that the model we used is fairly adequate.

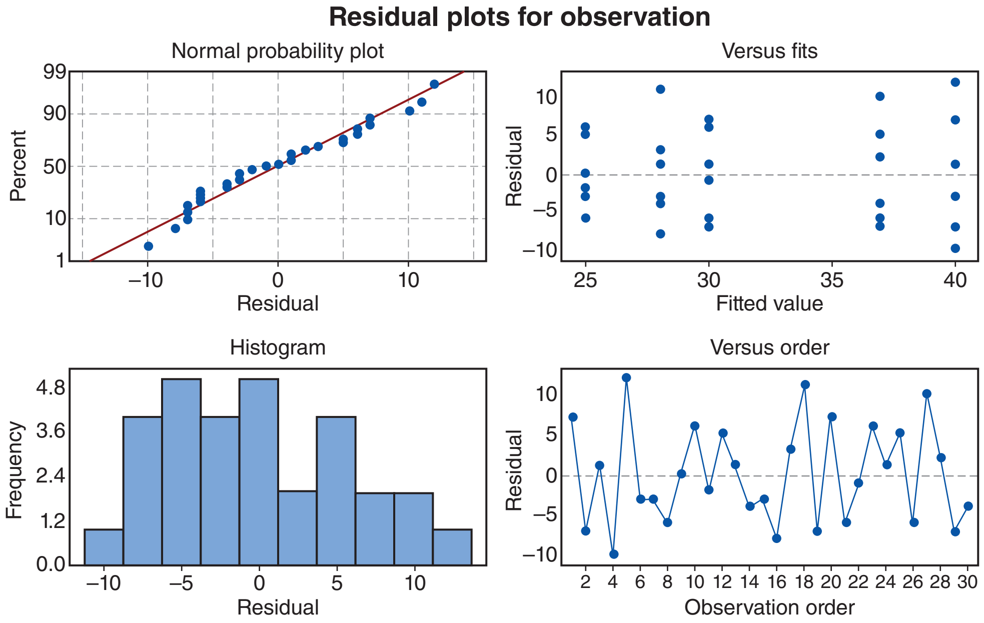


Figure 17.3.1 MINITAB residual analysis graphs.

17.3.3 Multiple Comparisons

In many industrial and scientific experiments, the hypothesis that there are no treatment differences is of little interest to the experimenter. The experimenter usually is interested in investigating which of the effects $\delta_1, \delta_2, \ldots, \delta_a$ have greatest influence on the response variable if the hypothesis that all effects are the same is rejected. As examples, an engineer needs to find the combination of metals that will produce wires with higher tensile strength, a physician may want to find which drug is more effective in treating a disease, or a chemist may want to find the amount of catalyst that gives the optimal yield of a chemical. The latter goal is usually achieved by testing hypotheses about various linear combinations of parameters of interest. In particular, the experimenter usually has an interest in hypotheses about contrasts among treatment effects.

Definition 17.3.1 A linear combination of the parameters $\mu_1, \mu_2, \ldots, \mu_a$ of the form $\sum_{i=1}^{a} c_i\mu_i$, where the constants c_i are such that $\sum_{i=1}^{a} c_i = 0$, is called a *contrast*.

For example, $\mu_1 - \mu_2, \mu_1 - 2\mu_2 + \mu_3, \mu_1 - \mu_2 - \mu_3 + \mu_4$ are contrasts. Contrasts of the form $\mu_i - \mu_j$ are called *simple contrasts* and are commonly used to compare two treatment effects.

From Theorem 17.2.2 given on the website, it follows that all contrasts $\sum_{j=1}^{a} c_j\delta_j$ in one-way layout experiments are estimable since

$$\eta = \sum_{j=1}^{a} c_j\mu_j = \sum_{j=1}^{a} c_j(\mu + \delta_j) = \sum_{j=1}^{a} c_j\delta_j \tag{17.3.34}$$

and (see after (17.3.8)) the δ_j are estimable.

Definition 17.3.2 Two contrasts $\eta_1 = \sum_{i=1}^{a} b_i\mu_i$, $\eta_2 = \sum_{i=1}^{a} c_i\mu_i$ are said to be orthogonal contrasts if $\sum_{i=1}^{a} b_i c_i = 0$.

For one-way ANOVA with a levels of a factor A, and under normality and independence of the y_{ij}s, orthogonal contrasts are independent in the sense that $(a-1)$ orthogonal contrasts divide the treatment sum of squares SS_A into $(a-1)$ independent components, each having one degree of freedom. That is, any conclusion made about one orthogonal contrast has no bearing on the conclusion made about any other orthogonal contrast.

Using Equation (17.3.34), we can easily see that the least-squares estimate of a contrast of the form $\eta = \sum_{j=1}^{a} c_j\mu_j$ is given by

$$\hat{\eta} = \sum_{j=1}^{a} c_j\hat{\mu}_j = \sum_{j=1}^{a} c_j\overline{y}_{.j}$$

This contrast is normally distributed with its estimated standard error

$$\hat{\sigma}_{\hat{\eta}} = \sqrt{\hat{\sigma}^2 \sum_{j=1}^{a} c_j^2/n_j} = \sqrt{MS_E \sum_{i=1}^{a} c_j^2/n_j}$$

Thus, we test a hypothesis (or find a confidence interval) about such a contrast at the α level of significance by employing a simple t-test. However, it is important to know that we cannot perform hypothesis testing at the α level of significance about several contrasts, say r, simultaneously by using a t-test at α level of significance for each test. That is, in order for the conclusions about the whole family of r hypotheses (or confidence intervals) to hold simultaneously, the significance value, or probability of type I error, should be equal to $1 - (1 - \alpha)^r$. For example, if $r = 2$ and $\alpha = 0.05$, then the probability of type I error for the conclusion about the two hypotheses to hold simultaneously is $1 - (0.95)^2 = 0.0975$.

Many methods are available in the literature to deal with simultaneous confidence intervals or testing simultaneously a set of r hypotheses. The more commonly used methods are due to Scheffe (1953), Tukey (1953), and Bonferroni (1936), which we discuss here. We note that the Dunnett (1964) method is usually used for the treatment versus control contrasts $\delta_i - \delta_1 (i = 2, \ldots, a)$, where treatment 1 is the control treatment. Scheffe and Tukey methods are commonly called the *S-method* and *T-method*, respectively. For more details, the reader can refer to Scheffe (1953, 1959) and Tukey (1953). The following definitions are useful in our discussion of the Scheffe and Tukey methods.

Definition 17.3.3 Let $\{\Psi_1, \ldots, \Psi_t\}$ be a set of linearly independent estimable functions, and let Ω be the set of all possible linear combinations $\sum_j c_j \psi_j$, where c_js are known constants. Then, the set Ω is called the *t-dimensional space* of *estimable functions.*

For example, in the model (17.3.2), we have $\{\psi_j\}$, $\psi_j = \delta_1 - \delta_j, i = 2, \ldots, a$ as a set of linearly independent estimable functions. The set Ω of all possible linear combinations of ψ_js forms an $(a - 1)$-dimensional space of estimable functions. Then, since ψ_j's are contrasts, every $\Psi \in \Omega$ is also a contrast.

Definition 17.3.4 Let ψ be an estimable function. Then, the estimate $\hat{\psi}$ is *not significantly different from zero* if the confidence interval (17.3.35), given below in Theorem 17.3.1, contains the point $\psi = 0$. Otherwise, $\hat{\psi}$ is significantly different from zero.

Definition 17.3.5 Let $\{Y_1, \ldots, Y_n\}$ be a random sample of size n from a population $N(\mu, \sigma^2)$ and let R be the sample range with $R = \max_i Y_i - \min_i Y_i$. Let S^2 be the mean-square estimator of σ^2 with m degrees of freedom. From Chapter 7, we know that S^2/σ^2 is distributed as χ^2_m, that is, as a Chi-square variable with m degrees of freedom (in (17.3.2), $m = N - a$). Then $R/\sqrt{S^2/m}$ is called the *Studentized range random variable*, and its distribution is called the *Studentized range distribution.*

Table A.13 provides the values of $q_{a,m;\ \alpha}$ the upper α percentage points of the *Studentized range distribution* q, where a is the number treatments being compared and m is the number of degrees of freedom for mean square error MS_E (mean-square estimator of σ^2).

The S-Method

Theorem 17.3.1 (Scheffe, 1953) *Under the model (17.3.2), the simultaneous probability for all $\psi \in \Omega$ to obey the inequality*

$$\hat{\psi} - \theta\hat{\sigma}_{\hat{\psi}} \le \psi \le \hat{\psi} + \theta\hat{\sigma}_{\hat{\psi}} \tag{17.3.35}$$

is $(1-\alpha)$ where θ is such that

$$\theta^2 = tF_{t,N-a;\ \alpha} \tag{17.3.36}$$

In (17.3.36), $t = a - 1$ if Ω is only the space of all contrasts. However, if we consider the larger space of all estimable functions, that is, the space generated by $\mu_1, \ldots, \mu_a$ where $\mu_j = \mu + \delta_j$ $(j = 1, \ldots, a)$, then $t = a$. Moreover, we can easily see that if $\psi = \sum_j c_j\delta_j$ is a contrast, then

$$\hat{\psi} = \sum_j c_j(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) = \sum_j c_j\bar{y}_{\cdot j}, \quad \text{since} \sum_j c_j = 0$$

Furthermore, we have for this case that

$$\sigma^2{}_{\hat{\psi}} = \sigma^2 \sum_j \left(\frac{c_j^2}{n_j}\right), \quad \hat{\sigma}^2{}_{\hat{\psi}} = MS_E \sum \left(\frac{c_j^2}{n_j}\right)$$

where n_j is the sample size of the jth sample (jth treatment). Note that for *multiple comparisons* in a one-way layout experiment, we use (17.3.36) in the confidence interval derived from (17.3.35) with $t = a - 1$.

Suppose now, referring to model (17.3.2), that the hypothesis $H_0 : \{\delta_1 = \delta_2 = \cdots = \delta_a\}$ is rejected at the α level of significance and we want to investigate which δs are different from each other. We consider all possible $a(a-1)/2$ simple contrasts $\{\psi_i\}$, that is, the difference of all the possible pairs of effects and obtain for each effect the interval $(\hat{\psi} \pm \theta\hat{\sigma}_{\hat{\psi}})$—see (17.3.35). We can then determine those ψs that are significantly different from zero at the α level, that is, which treatment effects are different from each other, by noting which interval does not contain 0.

The T-Method

Tukey's method, or the T-method, uses the distribution of the Studentized range statistic that we discussed earlier. The T-method is used, under certain restrictions, to study simultaneous confidence intervals about the contrasts among the parameters $\{\delta_1, \delta_2, \ldots, \delta_a\}$. The confidence intervals in the T-method are determined in terms of the unbiased estimates $\{\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_a\}$, and the upper α point of the Studentized range distribution under the condition that the $\hat{\delta}_j$s have equal variances. In the one-way layout, this condition of equal variances is met only if the sample sizes are equal. If the sample sizes are not equal, then the confidence coefficient for Tukey's multiple comparison method is known to be greater than $1 - \alpha$. Consequently the significance level is less than α. We state Tukey's result in the following theorem:

Theorem 17.3.2 *The probability that all possible contrasts $\psi = \sum_{j=1}^{a} c_j\delta_j$ simultaneously satisfy the inequalities*

$$\hat{\psi} - Ts\left(\frac{1}{2}\sum_{j=1}^{a}|c_j|\right) \le \psi \le \hat{\psi} + Ts\left(\frac{1}{2}\sum_{j=1}^{a}|c_j|\right) \tag{17.3.37}$$

is $1 - \alpha$ where $\hat{\psi} = \sum_{j=1}^{a} c_j\hat{\delta}_j, T = kq_{a,N-a;\alpha'}$, with k as a known constant (see below (17.3.39) and (17.3.40)), and $\sum c_j = 0$.

For the simple contrasts $\psi = \delta_i - \delta_j$, $i \neq j$, the inequalities (17.3.37) become

$$\hat{\psi} - Ts \le \psi \le \hat{\psi} + Ts \tag{17.3.38}$$

since $\dfrac{1}{2}\sum|c_j| = 1$ for simple contrasts.

Suppose now, referring to model (17.3.2), that $n_j = n$ for all j. Then the inequality (17.3.38) for any contrast $\psi = \sum_{j=1}^{a} c_j\delta_j$ becomes

$$\hat{\psi} - T\sqrt{MS_E}\left(\frac{1}{2}\sum_{j=1}^{a}|c_j|\right) \le \psi \le \hat{\psi} + T\sqrt{MS_E}\left(\frac{1}{2}\sum_{j=1}^{a}|c_j|\right) \tag{17.3.39}$$

where MS_E is the mean square error and $T = q_{a,N-a;\alpha}/\sqrt{n}$.

If the sample sizes are unequal, then inequality (17.3.39) holds with probability $\geq 1 - \alpha$, where

$$T = \frac{q_{a,N-a;\ \alpha}}{\sqrt{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \quad i \neq j. \tag{17.3.40}$$

The method with unequal sample sizes is sometimes called the *Tukey–Kramer method. Note*: The T-method for simple contrasts $\delta_i - \delta_j$, $i \neq j$, gives shorter intervals than the S-method. However, for general contrasts the situation is usually reversed.

The Bonferroni Method

The Bonferroni multiple comparison method is also quite popular with practitioners. We discuss that method briefly here. The Bonferroni method is quite similar to the ordinary t-test studied in Chapter 9, except that in the Bonferroni method we replace α with α/m, where m is the number of paired differences, contrasts, or linear combinations of model parameters we are studying. For example, if in a certain experiment, we are comparing five treatments and if we are interested in all pairwise comparisons, then m is the number of possible pairs, $m = \binom{5}{2} = 10$.

Theorem 17.3.3 *The probability that any set of m paired differences, contrasts, or linear combinations of model parameters ψ_i (i = 1, 2, . . . , m) simultaneously satisfy the inequalities*

$$\hat{\psi}_i - t_{N-a;\ \alpha/2m}\sqrt{MS_E}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} < \psi_i < \hat{\psi}_i + t_{N-a;\ \alpha/2m}\sqrt{MS_E}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{17.3.41}$$

is at least (1 - α).

The Bonferroni simultaneous confidence intervals for all pairs of treatment effects are usually wider than the Scheffe and Tukey confidence intervals.

Example 17.3.4 (Example 17.3.3 revisited) *Refer to the onion ring data in Example 17.3.3. We now illustrate the S-method, T-method, and Bonferroni method using these data. For convenience, we reproduce these data in Table 17.3.8.*

Solution: We now use the data in Table 17.3.8 to determine 95% simultaneous confidence intervals for all pairs of treatment effects using the Scheffe method, the Tukey method, and the Bonferroni method. For determining simultaneous confidence intervals for all pairs of treatment effects, we can use coded data.

Scheffe Method

From (17.3.35) and (17.3.36), simultaneous confidence intervals for all pairs of treatment effects $\delta_i - \delta_j$, $i \neq j$, are given by

$$\hat{\psi} \pm \theta\hat{\sigma}_{\hat{\psi}}$$

where θ is such that

$$\theta^2 = tF_{t,N-a;\ \alpha}$$

Table 17.3.8 Coded data $(y - 100)$ in grams of oil absorbed per batch.

Batch \ Oil	Olive	Soybean	Corn	Peanut	Sunflower
1	47	22	29	23	42
2	33	19	24	37	31
3	41	25	25	24	47
4	30	31	20	29	39
5	52	23	31	36	30
6	37	30	39	31	33
Totals	240	150	168	180	222

We know that $\delta_i - \delta_j,\ i \neq j$ are contrasts, $t = a - 1 = 4$, the F upper tail $\alpha = 0.05$ percentage point of the $F_{4,25}$-distribution is $F_{4,25;0.05} = 2.76$, $(N - a = 30 - 5 = 25)$, and from the ANOVA table in Example 17.3.3, $MS_E = 43.4$. Thus, using the Scheffe method, we have

$$\theta^2 = tF_{t,N-a;\ \alpha} = 4(2.76) = 11.04$$
$$\hat{\sigma}^2 \hat{\psi} = MS_E \sum_j \left(\frac{c_j^2}{n_j} \right) = 43.4 \left(\frac{1}{6} \right) ((1)^2 + (-1)^2) = 14.47$$

so that

$$\theta \hat{\sigma}_{\hat{\psi}} = \sqrt{11.04} \times \sqrt{14.47} = 12.64.$$

For example, the confidence interval for the contrast $\delta_1 - \delta_2$ is given by

$$(\hat{\psi} \pm \theta \hat{\sigma}_{\hat{\psi}}) = (15 \pm 12.64) = (2.36, 27.64)^*$$

since $\hat{\psi} = \hat{\delta}_1 - \hat{\delta}_2 = 40 - 25 = 15$.
Similarly, the confidence intervals for all other contrasts $\delta_i - \delta_j,\ i \neq j$ are

$(-0.64,\ 24.64)$	for $\delta_1 - \delta_3$
$(-2.64,\ 22.64)$	for $\delta_1 - \delta_4$
$(-9.64,\ 15.64)$	for $\delta_1 - \delta_5$
$(-15.64,\ 9.64)$	for $\delta_2 - \delta_3$
$(-17.64,\ 7.64)$	for $\delta_2 - \delta_4$
$(-24.64,\ 0.64)$	for $\delta_2 - \delta_5$
$(-14.64,\ 10.64)$	for $\delta_3 - \delta_4$
$(-21.64,\ 3.64)$	for $\delta_3 - \delta_5$
$(-19.64,\ 5.64)$	for $\delta_4 - \delta_5$

The starred confidence intervals are those that do not contain the zero point. Thus, the only confidence interval that does not contain 0 is the contrast $\delta_1 - \delta_2$. Using the Scheffe method, we conclude that the only treatment effects that are significantly different from each other at the 5% level of significance are treatments 1 and 2. In our discussion, the confidence intervals with such significant differences are highlighted by $()^*$ symbol

Tukey Method

As noted earlier, for the simple contrasts $\psi = \delta_i - \delta_j$, $i \neq j$, the Tukey inequalities are given by

$$\hat{\psi} - Ts \leq \psi \leq \hat{\psi} + Ts$$

with $T = q_{a,N-a;\alpha}/\sqrt{n}$ if sample sizes are equal ($n_i = n_j = n$), or

$$\frac{1}{\sqrt{2}} q_{a,N-a;\ \alpha} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}, \quad i \neq j$$

if sample sizes are unequal. In the present example, sample sizes are equal with $n_i = n_j = 6$, $a = 5$, $N - a = 25$, $\alpha = 0.05$, $S^2 = MS_E = 43.4$ and from Table A.13, $q_{5,25;0.05} \approx 4.15$. Then, the confidence interval for the contrast $\delta_1 - \delta_2$ is given by

$$\left(15 \pm \frac{1}{\sqrt{6}}(4.15)\sqrt{43.4}\right) = (15 \pm 11.16) = (3.84, 26.16)^*$$

Similarly, the confidence intervals for all other contrasts $\delta_i - \delta_j$, $i \neq j$, are

(0.84, 23.16)*	for $\delta_1 - \delta_3$
(−1.16, 21.16)	for $\delta_1 - \delta_4$
(−8.16, 14.16)	for $\delta_1 - \delta_5$
(−14.16, 8.16)	for $\delta_2 - \delta_3$
(−16.16, 6.16)	for $\delta_2 - \delta_4$
(−23.16,−0.84)*	for $\delta_2 - \delta_5$
(−13.16, 9.16)	for $\delta_3 - \delta_4$
(−20.16, 2.16)	for $\delta_3 - \delta_5$
(−18.16, 4.16)	for $\delta_4 - \delta_5$

Tukey's confidence intervals for all contrasts $\delta_i - \delta_j$, $i \neq j$ are narrower than Scheffe's confidence intervals. Moreover, in Tukey's confidence intervals, δ_1 is significantly different from δ_2 and δ_3; also, δ_2 and δ_5 are significantly different from each other. These confidence intervals match those obtained using MINITAB in Example 17.3.3.

Bonferroni Method

We saw earlier that for the simple contrasts $\psi_j = \delta_1 - \delta_j, j = 2, 3, \ldots, a$, the Bonferroni inequalities are given by

$$\hat{\psi}_j - t_{N-a;\alpha/2m}\sqrt{MS_E}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_j}\right)} < \psi_j < \hat{\psi}_j + t_{N-a;\alpha/2m}\sqrt{MS_E}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_j}\right)}$$

In the present example, sample sizes are equal, we have five treatments, and the number of paired contrasts is $m = \binom{5}{2} = 10$. Thus, for example, the confidence interval for the contrast $\delta_1 - \delta_2$ is given by

$$\left(15 \pm t_{25;\ 0.0025}\sqrt{43.4}\sqrt{\left(\frac{1}{6}+\frac{1}{6}\right)}\right) = (15 \pm 11.71) = (3.29, 26.71)^*$$

Similarly, the confidence intervals for all other contrasts $\delta_i - \delta_j, i \neq j$ are

(0.29, 23.71)*	for $\delta_1 - \delta_3$
(−1.71, 21.71)	for $\delta_1 - \delta_4$
(−8.71, 14.71)	for $\delta_1 - \delta_5$
(−14.71, 8.71)	for $\delta_2 - \delta_3$
(−16.71, 6.71)	for $\delta_2 - \delta_4$
(−23.71,−0.29)*	for $\delta_2 - \delta_5$
(−13.71, 9.71)	for $\delta_3 - \delta_4$
(−20.71, 2.71)	for $\delta_3 - \delta_5$
(−18.71, 4.71)	for $\delta_4 - \delta_5$

Note that Bonferroni's confidence intervals for all contrasts $\delta_i - \delta_j$, $i \neq j$ are narrower than Scheffe's confidence intervals but wider than Tukey's confidence intervals. Moreover, using Bonferroni's confidence intervals, we arrive at the same conclusion as by using Tukey's confidence intervals, that is, that δ_1 is significantly different from δ_2 and δ_3, and δ_2 and δ_5 are significantly different from each other.

17.3.4 Determination of Sample Size

In order to run any experiment, the experimenter must ascertain how many times to replicate each treatment. Here, we discuss a technique usually called the *confidence interval estimation technique* and assume that all treatments are replicated the same number of times; that is, the sample sizes are equal. For some other techniques, we refer the reader to Montgomery (2009).

The confidence interval estimation technique assumes that the experimenter has some experience dictating how wide he/she wants these confidence intervals to be. For instance, in the onion ring example, suppose that the experimenter would like the $(1 - \alpha)$ confidence interval to be ±10 g accurate for each pair of treatments. In this example, $n_j = n = 6, j = 1, 2, \ldots, a; a = 5$. From our earlier discussion in this chapter, we know that the accuracy of the confidence interval is given by

$$\pm t_{N-a;\alpha/2}\sqrt{\frac{2MS_E}{n}} \tag{17.3.42}$$

where from some earlier experience we know that $\sqrt{MS_E} = S$ is approximately 7 units. Suppose now that $\alpha = 0.05$. Using (17.3.42), we can determine the accuracy of the confidence interval for different values of n. Then, the desirable sample size is the one that produces accuracy of less than or equal to ±10 g. In this example, we obtain $n = 5$ approximately, since $N - a = a(n-1) = 5(n-1)$, and if $Q(n)$ denotes the accuracy of the confidence interval, then

$$Q(n) = t_{5(n-1);0.025}\sqrt{\frac{2MS_E}{n}} = 7t_{5(n-1);0.025}\sqrt{\frac{2}{n}}$$

Thus, $Q(5) = 7(2.086) \times \sqrt{0.4} = 9.235 \leq 10$ and $n = 5$ is the largest value of n for which $Q(n) \leq 10$.

17.3.5 The Kruskal–Wallis Test for One-Way Layouts (Nonparametric Method)

In practice, sometimes the normality assumption in the model (17.3.2) does not hold. In such situations, we need to analyze our data by an alternative procedure, that is, a distribution-free or nonparametric procedure. One such procedure has been developed by Kruskal and Wallis (1952). This test is used to test the hypothesis

$$H_0: \ \delta_1 = \delta_2 = \cdots = \delta_a \text{ versus } \quad H_1: \text{ At least one } \delta \text{ is different}$$

The Kruskal–Wallis test uses the following procedure:

1. Write all the observations y_{ij} in the ascending order and rank them from 1 to N, $N = \sum_{j=1}^{a} n_j$, assigning the rank 1 to the smallest observation and the rank N to the largest observation. If some observations have the same value, then they have tied ranks. To break these ties, we assign the average rank to each tied observation.
2. Replace each observation y_{ij} by its rank, say r_{ij}.
3. Then the Kruskal–Wallis test statistic is given by

$$H = \frac{1}{S^2}\left(\sum_{j=1}^{a}\frac{T_{\cdot j}^2}{n_j} - \frac{N(N+1)^2}{4}\right) \tag{17.3.43}$$

where n_j is the number of observations in the jth treatment and $T_{\cdot j} = \sum_{i=1}^{n_j} r_{ij}$ is the sum of the ranks of the observations in the jth treatment. Also $N = \sum_{j=1}^{a} n_j$, and S^2 is the variance of the ranks and is given by

$$\begin{aligned} S^2 &= \tfrac{1}{N-1}\left(\sum_{j=1}^{a}\sum_{i=1}^{n_j} r_{ij}^2 - \tfrac{N(N+1)^2}{4}\right) = \tfrac{1}{N-1}\left(\sum_{t=1}^{N} t^2 - \tfrac{N(N+1)^2}{4}\right) \\ &= \tfrac{1}{N-1}\left(\tfrac{N(N+1)(2N+1)}{6} - \tfrac{N(N+1)^2}{4}\right) = \tfrac{N(N+1)}{12} \end{aligned} \tag{17.3.44}$$

If there are no ties or the number of ties is limited, then it can be seen that the test statistic is given by

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{a}\frac{T_{\cdot j}^2}{n_j} - 3(N+1) \tag{17.3.45}$$

4. The test statistic H is approximately distributed as χ^2_{a-1}. Thus the null hypothesis is rejected at α level of significance whenever

$$H > \chi^2_{a-1;\ \alpha} \tag{17.3.46}$$

Example 17.3.5 (Onion rings, revisited) *The onion ring data and their corresponding ranks in parentheses are shown in Table 17.3.9.*

Table 17.3.9 Coded data $(y - 100)$ in grams of oil absorbed per batch and their ranks (in parentheses).

Oil Batch	Olive	Soybean	Corn	Peanut	Sunflower
1	47 (28.5)	22 (3)	29 (10.5)	23 (4.5)	42 (27)
2	33 (19.5)	19 (1)	24 (6.5)	37 (22.5)	31 (16.5)
3	41 (26)	25 (8.5)	25 (8.5)	24 (6.5)	47 (28.5)
4	30 (13)	31 (16.5)	20 (2)	29 (10.5)	39 (24.5)
5	52 (30)	23 (4.5)	31 (16.5)	36 (21)	30 (13)
6	37 (22.5)	30 (13)	39 (24.5)	31 (16.5)	33 (19.5)
Rank totals	139.5	46.5	68.5	81.5	129

We now use the test statistic defined by (17.3.43) and (17.3.44). Thus we have

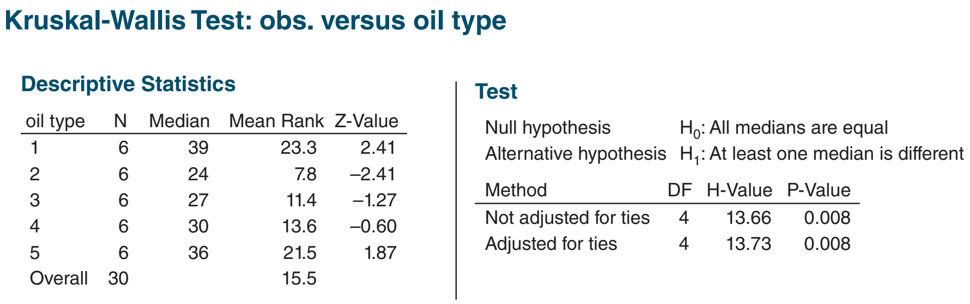
$$S^2 = \frac{30(31)}{12} = 77.5 \Rightarrow H = \frac{1}{77.5}(8266.33 - 7207.5) = 13.66$$

Since the observed $H = 13.66$, which is greater than $\chi^2_{4;0.05} = 9.49$, we reject the null hypothesis and conclude the treatment effects are not equal. Clearly, this conclusion is the same as reached in Example 17.3.3.

Example 17.3.6 (Using MINITAB and R and applying the Kruskal–Wallis test.) *Analyze the onion ring data in Example 17.3.3, using the Kruskal–Wallis test.*

MINITAB

1. Enter the data in column C1.
2. Enter the treatment identifiers in column C2. We name these columns obs. and oil type, respectively.
3. Select **Stat** > **Nonparametrics** > **Kruskal–Wallis ....**
4. In the dialog box that appears, type obs. in the box next to **Response** and enter oil type in the box next to **Factor**. Then click **OK**. The MINITAB output appears in the session window as shown below.



```
Kruskal-Wallis Test: obs. versus oil type

Descriptive Statistics                        Test

oil type   N  Median  Mean Rank  Z-Value      Null hypothesis         H0: All medians are equal
1          6      39       23.3     2.41      Alternative hypothesis  H1: At least one median is different
2          6      24        7.8    –2.41
3          6      27       11.4    –1.27      Method                DF  H-Value  P-Value
4          6      30       13.6    –0.60      Not adjusted for ties  4    13.66    0.008
5          6      36       21.5     1.87      Adjusted for ties      4    13.73    0.008
Overall   30               15.5
```

Since the p-value is 0.008, which is smaller than $\alpha = 0.05$, we reject the null hypothesis that all treatments have the same effect.

USING R

Solution: The R function ‘kruskal.test()’ can be used to conduct the required Kruskal–Wallis test as shown in the following R-code.

```
obs = c(47,33,41,30,52,37,22,19,25,31,23,30,29,24,25,20,31,39,23,37,24,29,36,31,42,
31,47,39,30,33)
oil.type = c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,5,5,5,5,5,5)

#Kruskal-Wallis test
kruskal.test(obs ~ oil.type)

#R output
Kruskal-Wallis rank sum test
data: obs by oil.type
Kruskal-Wallis chi-squared = 13.73, df = 4, p-value = 0.00821
```

PRACTICE PROBLEMS FOR SECTION 17.3

1. Determinations were made of the production of a chemical using four catalytic methods I, II, III, and IV, with the results shown below:

Catalytic method	I	II	III	IV
	45.4	50.7	48.7	52.7
	47.6	49.6	47.6	54.1
Yield	46.3	48.8	45.7	53.2
	44.5		46.9	51.5
				50.9

Test the hypothesis H_0, that the effects due to the catalytic methods are the same. Use $\alpha = 0.05$.

2. Refer to Problem 1. (a) Estimate all the effects. (b) Use the S-method for a multiple pairwise comparisons test on effects of all catalytic methods.
3. Three varieties of wheat were tested for productivity by employing a completely randomized experiment. The experiment is replicated four times, the yields are coded, and the coded values (yield − 100) are as presented below.
   (a) Construct the ANOVA table and test the null hypothesis that the varieties are equally productive at the 5% level of significance.
   (b) If the hypothesis H_0 is rejected, estimate the mean effects of the three varieties:

Variety	I	II	III
	43.7	47.6	43.3
Yield	39.6	45.9	42.9
	41.0	43.0	42.0
	42.3	44.5	41.7

4. An experiment is carried out to compare five brands of gasoline. The following data give five observations on octane numbers of each of five gasolines:

Gasoline brands	1	2	3	4	5
	77	71	73	75	77
	72	73	67	72	73
Octane	79	73	71	72	72
	79	77	69	69	76
	76	67	70	73	70

   (a) Construct ANOVA table for these data.
   (b) Use the Tukey method to perform pairwise multiple comparisons test for all five gasolines. Use $\alpha = 0.05$
5. Five different types of training (T_1, T_2, T_3, T_4, T_5) were given to 23 technicians. The allocation of training to technicians was random. Their performance on a specific project was evaluated. Scores in coded data are given below:

T_1	T_2	T_3	T_4	T_5
8	6	10	6	10
7	10	9	8	9
5	8	9	7	5
6	9	7	7	
	9	10	8	
	7			

   (a) Construct an ANOVA table for these data.

(b) Use the Tukey method to perform a pairwise multiple comparisons test among the five training techniques. Use $\alpha = 0.10$.

6. Refer to the data in Problem 4. Use the Kruskal–Wallis test statistic to test a hypothesis that there is no significant difference of octane numbers in the five gasolines. Use $\alpha = 0.05$.
7. Refer to Problem 1 above. Use the Bonferroni method to perform a pairwise multiple comparisons test on effects of the four catalytic methods. Use $\alpha = 0.05$.
8. Refer to Problem 1 above. Use the Kruskal–Wallis test statistic to test the hypothesis that there is no significant difference between the catalytic methods. Use $\alpha = 0.05$.
9. Refer to Problem 3 above. Use the Tukey test to perform a pairwise multiple comparisons test for all three varieties. Use $\alpha = 0.01$.
10. Refer to Problem 3 above. Use the Kruskal–Wallis test statistic to test a hypothesis that there is no significant difference between the varieties. Use $\alpha = 0.01$.
11. Refer to Problem 3 above. Use the S-method to perform a pairwise multiple comparisons test for all three varieties and draw your conclusions. Do you get the same conclusions as in Problems 9 and 10 above? Use $\alpha = 0.01$.
12. Refer to Problem 5 above. Use the Kruskal–Wallis test statistic to test a hypothesis that there is no significant difference in training techniques. Use $\alpha = 0.01$.
13. Refer to Problem 5 above. Use the Bonferroni method to perform a pairwise multiple comparisons test for all training techniques and draw your conclusions. Do you get the same conclusions as in Problems 5 and 12 above? Use $\alpha = 0.05$.

17.4 RANDOMIZED COMPLETE BLOCK (RCB) DESIGNS

In the distribution of one-way layout experiments in Section 17.3, we assumed that all the experimental units are homogeneous with respect to all known sources of variations. In such experiments, treatments are randomly assigned to the experimental units. Hence, these designs are also called *completely randomized designs.*

In practice, however, it may be difficult to find a large number of experimental units that are completely homogeneous with respect to all known sources of variations. For example, in an experiment where interest lies in comparing the effects of various doses of a drug, we may need a large number of patients who are being treated with that drug. In such an experiment, it may not be possible to have a large number of patients who have the same condition with respect to all factors, namely the advance of the disease, family history, weight, age, and sex.

In these situations, when all the experimental units are not homogeneous, we adopt a technique called *blocking.* That is, we divide the experimental units into various groups, say "b" groups of equal size, say a, such that within each subgroup the experimental units are as homogeneous as possible with respect to all known sources of variations. Then, within each subgroup, the a treatments are randomly assigned to the a experimental units. Such experiments are called *randomized complete block* (*RCB*) *designs.* They are complete in the sense that within each subgroup we have a complete replication of all the treatments, applied randomly.

These designs have widespread applications in many industrial and other experiments, for example, testing tensile strength of wires produced using different machines, testing different methods of production using various operators, testing different brands of tires for different passenger cars, testing different teaching methods, or testing a certain number of drugs on a group of animals. In these examples, the different blocks consist of machines, operators, cars, students, and animals, respectively.

The model for a RCB design where the number of treatments is a and the number of blocks b is given by

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b \tag{17.4.1}$$

where y_{ij} is an observation generated using the ith treatment in the jth block, μ is the general mean, α_i is the effect of the ith treatment, β_j is the effect of the jth block, and ε_{ij} is a random error. We assume that the ε_{ij} are independently and identically distributed as $N(0, \sigma^2)$. Furthermore, we assume that the α_i and β_j are such that they satisfy the *side conditions*

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0 \tag{17.4.2}$$

and that $\mu,\ \alpha_i,\ i = 1, 2, \ldots, a,\ \beta_j,\ j = 1, 2, \ldots, b$ are unknown constants.

The hypotheses of principal interest in this model are

1. $H_0:\ \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ versus $H_1:$ not all $\alpha_i = 0$
2. $H_0:\ \beta_1 = \beta_2 = \cdots = \beta_b = 0$ versus $H_1:$ not all $\beta_j = 0$

Note that the two null hypotheses H_0 indicate that all treatment effects are equal and all block effects are equal. The data obtained from a RCB experiment are laid out in Table 17.4.1, where $T_{i\cdot} = \sum_{j=1}^{b} y_{ij}$, $T_{\cdot j} = \sum_{i=1}^{a} y_{ij}$, $G = \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij} = \sum_{i=1}^{a} T_{i\cdot} = \sum_{j=1}^{b} T_{\cdot j} = y_{\cdot\cdot}$, and $\bar{y}_{\cdot j} = T_{\cdot j}/a$ is the average of the observations in the jth block, $\bar{y}_{i\cdot} = T_{i\cdot}/b$ is the average of the observations generated by the use of the ith treatment, and $\bar{y}_{\cdot\cdot} = G/ab$ is the grand average of all the observations.

The error sum of squares under the model (17.4.1) is given by

$$Q = \sum_{i} \sum_{j} (y_{ij} - \mu - \alpha_i - \beta_j)^2 \tag{17.4.3}$$

The least-squares estimates of $\mu,\ \alpha_i, i = 1, 2, \ldots, a, \beta_j, j = 1, 2, \ldots, b$ are obtained by minimizing Q in (17.4.3), subject to the conditions (17.4.2). This is accomplished by

Table 17.4.1 Randomized complete block design layout.

Blocks Treatments	B_1	B_2	B_3	$\cdots$	B_j	$\cdots$	B_b	Totals
A_1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1j}	$\cdots$	y_{1b}	$T_{1\cdot}$
A_2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2j}	$\cdots$	y_{2b}	$T_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_a	y_{a1}	y_{a2}	y_{a3}	$\cdots$	y_{aj}	$\cdots$	y_{ab}	$T_{a\cdot}$
Totals	$T_{\cdot 1}$	$T_{\cdot 2}$	$T_{\cdot 3}$	$\cdots$	$T_{\cdot j}$	$\cdots$	$T_{\cdot b}$	G

equating to zero the partial derivatives of Q with respect to μ, α_i, β_j, and solving these equations subject to the constraints (17.4.2). Denoting the solutions by $(\hat{\mu},\ \hat{\alpha}_i,\ \hat{\beta}_j)$, we obtain

$$\begin{cases} \sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0 \\ \sum\limits_{j=1}^{b}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0, \quad i = 1, 2, \ldots, a \\ \sum\limits_{i=1}^{a}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0, \quad j = 1, 2, \ldots, b \\ \sum\limits_{i=1}^{a}\hat{\alpha}_i = 0, \quad \sum\limits_{i=1}^{b}\hat{\beta}_j = 0 \end{cases} \tag{17.4.4}$$

Solving the system of Equations (17.4.4), we find that

$$\hat{\mu} = \overline{y}_{..}, \quad \hat{\alpha}_i = \overline{y}_{i.} - \overline{y}_{..}, \quad \hat{\beta}_j = \overline{y}_{.j} - \overline{y}_{..} \tag{17.4.5}$$

where $\overline{y}_{..} = \sum\sum y_{ij}/N,\ N = ab,\ \overline{y}_{i.} = \sum_{j=1}^{b} y_{ij}/b$, and $\overline{y}_{.j} = \sum_{i=1}^{a} y_{ij}/a$.

The total variation sum of squares is

$$SS_{tot} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2$$

We can split the total sum of squares into various components, that is, sum of squares due to error, treatments, and blocks. So we write

$$\begin{aligned} SS_{tot} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2 &= \textstyle\sum_{i=1}^{a}\sum_{j=1}^{b}((y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} - \overline{y}_{..}) + (\overline{y}_{i.} - \overline{y}_{..}) + (\overline{y}_{.j} - \overline{y}_{..}))^2 \\ &= \textstyle\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2 + b\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2 \\ &\quad + a\textstyle\sum_{j=1}^{b}(\overline{y}_{.j} - \overline{y}_{..})^2 \end{aligned} \tag{17.4.6}$$

since all cross-product terms vanish. We now use the notation

$$SS_{treat} = b\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2 \tag{17.4.7}$$

$$SS_{bl} = a\sum_{j=1}^{b}(\overline{y}_{.j} - \overline{y}_{..})^2 \tag{17.4.8}$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2 \tag{17.4.9}$$

It can be shown that *under normality*, SS_E, SS_{treat}, and SS_{bl} are *independently* distributed as $\sigma^2\chi^2_{(a-1)(b-1)}$, $\sigma^2\chi^2_{(a-1)}(\lambda_1)$ and $\sigma^2\chi^2_{(b-1)}(\lambda_2)$, respectively, where λ_1 and λ_2 are the parameters of noncentrality and are given by $\lambda_1 = b\sum_i\alpha_i^2/\sigma^2$ and $\lambda_2 = a\sum_j\beta_j^2/\sigma^2$.

Note, however, that under the null hypotheses, the corresponding noncentrality parameters are zero; that is, SS_{treat} and SS_{bl} are distributed as central Chi-squares whereas SS_E, regardless of any hypothesis, is always distributed as a central chi-square. Thus, under the null hypothesis, we have

$$\frac{SS_{treat}/(a-1)}{SS_E/(a-1)(b-1)} = \frac{MS_{treat}}{MS_E} \sim F_{(a-1),(a-1)(b-1)} \tag{17.4.10}$$

and

$$\frac{SS_{bl}/(b-1)}{SS_E/(a-1)(b-1)} = \frac{MS_{bl}}{MS_E} \sim F_{(b-1),(a-1)(b-1)} \tag{17.4.11}$$

Hence, we reject the hypothesis that all the treatments have the same effect, that is, all $\alpha_i = 0$ at the α level of significance if

$$\frac{MS_{treat}}{MS_E} \geq F_{(a-1),(a-1)(b-1);\alpha} \tag{17.4.12}$$

Similarly, we reject the hypothesis that all blocks have the same effect, that is, all $\beta_j = 0$ at the α level of significance if

$$\frac{MS_{bl}}{MS_E} \geq F_{(b-1),(a-1)(b-1);\alpha} \tag{17.4.13}$$

where $MS_{treat} = SS_{treat}/(a-1), MS_{bl} = SS_{bl}/(b-1),$ and $MS_E = SS_E/[(a-1)(b-1)]$. We summarize these results in Table 17.4.2.

From Table 17.4.2, it is obvious that an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = MS_E = \frac{1}{(a-1)(b-l)} \sum\sum (Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{\cdot\cdot})^2 \tag{17.4.14}$$

no matter whether $\alpha_i = 0, i = 1, 2, \ldots, a,$ and/or $\beta_j = 0, j = 1, 2, \ldots, b,$ is true or false. The calculations of the sums of squares for the analysis of variance above can be obtained more conveniently as follows:

$$SS_{treat} = \sum_i \frac{T_{i\cdot}^2}{b} - \frac{G^2}{ab} \tag{17.4.15}$$

Table 17.4.2 ANOVA table for an RCB-design.

Source of variation	*DF*	Sum of squares (*SS*)	*MS*	*E*(*MS*)	*F*-ratio
Treatments	$a-1$	$SS_{treat} = b\sum_{i=1}^{a}(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	MS_{treat}	$\sigma^2 + \frac{b\sum\alpha_i^2}{\alpha-1}$	$\frac{MS_{treat}}{MS_E}$
Blocks	$b-1$	$SS_{bl} = a\sum_{j=1}^{b}(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2$	MS_{bl}	$\sigma^2 + \frac{a\sum\beta_j^2}{b-1}$	$\frac{MS_{bl}}{MS_E}$
Error	$(a-1)(b-1)$	$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y}_{\cdot\cdot})^2$	MS_E	σ^2	
Total	$ab-1$	$\sum\sum(y_{ij} - \overline{y}_{\cdot\cdot})^2$			

$$SS_{bl} = \sum_j \frac{T_{\cdot j}^2}{a} - \frac{G^2}{ab} \tag{17.4.16}$$

$$SS_{tot} = \sum_i \sum_j y_{ij}^2 - \frac{G^2}{ab} \tag{17.4.17}$$

$$SS_E = SS_{tot} - SS_{bl} - SS_{treat} = \sum_i \sum_j y_{ij}^2 - \sum_i \frac{T_{i\cdot}^2}{b} - \sum_j \frac{T_{\cdot j}^2}{a} + \frac{G^2}{ab} \tag{17.4.18}$$

We need to point out that SS_E is a typical "error" sum of squares, for each term is the square of the residual of y_{ij} after model (17.4.1) has been fitted. Indeed, the fit of the model is

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}) = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$$

so that the residual of y_{ij} is

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{y}_{i\cdot} - \hat{y}_{\cdot j} + \hat{y}_{\cdot\cdot}$$

Examining (17.4.9), we see that

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} e_{ij}^2$$

Example 17.4.1 (Blood pressure) *Suppose that 24 patients who have high blood pressure are being treated with six different medications. Since it was not possible to find a large group of patients having similar risk factors such as, age, race, weight, family history, gender, and comorbidities, the design selected to test these medications was a RCB design. All the patients were divided into four blocks, each block containing six patients, with patients having similar risk factors included in the same block. Then, within each block, the six medications were randomly administered to the six patients. After the patients were treated for two months, their blood pressures were checked. Table 17.4.3 shows by how much the patient's systolic blood pressure reading decreased. We assume that the various assumptions of the RCB design model were valid for these clinical trials.*

Table 17.4.3 Data on 24 patients being treated for high blood pressure.

Blocks		1	2	3	4	Totals
	1	32	36	29	33	130
	2	18	28	21	30	97
Treatments (medicines)	3	22	23	28	24	97
	4	19	32	37	26	114
	5	24	24	25	21	94
	6	27	20	30	26	103
Totals		142	163	170	160	635

Table 17.4.4 ANOVA table for the data in Table 17.4.3.

Source	*DF*	*SS*	*MS*	*F*-ratio
Treatments (medicines)	5	238.71	47.74	2.28
Blocks	3	71.13	23.71	1.13
Error	15	314.12	20.94	
Total	23	623.96		

Using Equations (17.4.15)–(17.4.18) for calculating various sums of squares for the data in Table 17.4.3, we have

$$SS_{treat} = \frac{(130)^2 + (97)^2 + \cdots + (103)^2}{4} - \frac{(635)^2}{24} = 17,039.75 - 16,801.04 = 238.71$$

$$SS_b = \frac{(142)^2 + (163)^2 + (170)^2 + (160)^2}{6} - \frac{(635)^2}{24} = 16,872.17 - 16,801.04 = 71.13$$

$$SS_{tot} = (32)^2 + \cdots + (26)^2 - \frac{(635)^2}{24} = 17{,}425 - 16,801.04 = 623.96$$

$$SS_E = SS_{tot} - SS_{treat} - SS_{bl} = 623.96 - 238.71 - 71.13 = 314.12$$

We summarize the results above in Table 17.4.4 (DF = degrees of freedom).

To test the hypothesis that effects due to different medicines are the same, we find that the observed $F = MS_{treat}/MS_E$ is

$$F = \frac{47.74}{20.94} = 2.28 < F_{5,15;0.05} = 2.9013$$

We can conclude that the effects of the medications are not significantly different. (Similarly, the reader should verify that the block effects are not significantly different.)

Notes:

1. In an RCB-design, blocks are introduced in order to eliminate the effects of one nuisance variable. For instance, in Example 17.4.1, the difference between the patients is a nuisance variable.
2. In an RCB-design, the experimental units within each block are homogeneous, but between blocks, they are heterogeneous.
3. If, in an RCB-design, any of the hypotheses of equal effects is rejected, then we can use the S-method, the T-method, or the Bonferroni method to find simultaneous confidence intervals for all pairs of treatment effect or of block effects. All calculations are done in a similar manner as for completely randomized designs, except that the MS_E factor is the one given in Table 17.4.2 with $(a-1)(b-1)$ degrees of freedom.
4. In an RCB-design, it is usually the case that the hypotheses of treatment effects are of prime interest.

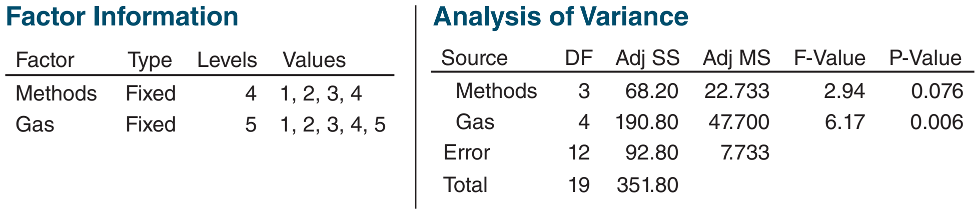
Example 17.4.2 (Gasoline octane levels) *The quality of gasoline is usually determined by its octane number. An experimenter determines the octane numbers of five gasolines using four different methods. Since "Methods" is a nuisance variable, the experimenter decided to use an RCB-design. The experiment in each block was carried out in random order. The data obtained are shown in Table 17.4.5. Analyze these data using MINITAB and R.*

Table 17.4.5 Octane data for five gasolines.

Blocks (methods)		1	2	3	4	Totals
	1	92	90	96	93	371
	2	88	87	84	80	339
Treatments (gasolines)	3	89	87	85	84	345
	4	90	89	82	86	347
	5	87	88	80	81	336
Totals		446	441	427	424	1738

MINITAB

1. Enter the data in column C1.
2. Enter the block (Methods) identifiers (1, 2, 3, 4) in column C2.
3. Enter the treatments (Gasolines) identifiers (1, 2, 3, 4, 5) in column C3. We name these columns as Obs., Methods, and Gas, respectively.
4. Select **Stat** > **Anova** > **General Linear Model** > **Fit General Linear Model:**.
5. In the dialog box that appears type in Obs. in the box below **Response**, Methods and Gas in the box below **Factors**. Then click **OK**. The MINITAB output appears in the Session window as shown below:



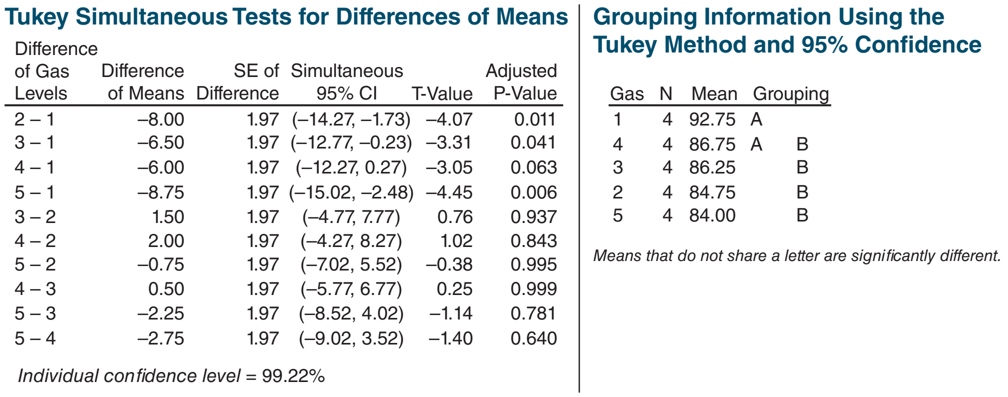
Factor Information

Factor	Type	Levels	Values
Methods	Fixed	4	1, 2, 3, 4
Gas	Fixed	5	1, 2, 3, 4, 5

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Methods	3	68.20	22.733	2.94	0.076
Gas	4	190.80	47.700	6.17	0.006
Error	12	92.80	7.733		
Total	19	351.80			

Since the p-value for gas is 0.006, we reject the null hypothesis that all gasolines have equal effects at 5% level of significance. However, the hypothesis that blocks have no effect is not rejected because the p-value is greater than 5%. To conduct the multiple comparisons, select **Stat** > **Anova** > **General Linear Model** > **Comparisons:**. Then select the comparison **Method** (e.g., Tukey) and **Choose terms for comparisons** (e.g., Gas) from the new window appears. Also, if needed we can select additional Options, Graphs, etc. The MINITAB output appears in the Session window as shown below:



Tukey Simultaneous Tests for Differences of Means

Difference of Gas Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
2 – 1	–8.00	1.97	(–14.27, –1.73)	–4.07	0.011
3 – 1	–6.50	1.97	(–12.77, –0.23)	–3.31	0.041
4 – 1	–6.00	1.97	(–12.27, 0.27)	–3.05	0.063
5 – 1	–8.75	1.97	(–15.02, –2.48)	–4.45	0.006
3 – 2	1.50	1.97	(–4.77, 7.77)	0.76	0.937
4 – 2	2.00	1.97	(–4.27, 8.27)	1.02	0.843
5 – 2	–0.75	1.97	(–7.02, 5.52)	–0.38	0.995
4 – 3	0.50	1.97	(–5.77, 6.77)	0.25	0.999
5 – 3	–2.25	1.97	(–8.52, 4.02)	–1.14	0.781
5 – 4	–2.75	1.97	(–9.02, 3.52)	–1.40	0.640

Individual confidence level = 99.22%

Grouping Information Using the Tukey Method and 95% Confidence

Gas	N	Mean	Grouping	
1	4	92.75	A	
4	4	86.75	A	B
3	4	86.25		B
2	4	84.75		B
5	4	84.00		B

Means that do not share a letter are significantly different.

Using these comparisons, we can easily see which means are different if the null hypothesis of equal means is rejected. For example, examining confidence intervals and grouping information in the above table, the mean of gasoline 1 is different from that of all other gasolines except gasoline 4, since only the confidence interval for Gas 4–Gas 1 does contain the zero and the other confidence intervals (Gas 2–Gas 1, Gas 3–Gas 1, and Gas 5–Gas 1) does not contain the zero.

USING R

Solution: The R function ‘aov()’ can be used to fit the required RCB design as shown in the following R-code.

```
Obs = c(92,88,89,90,87,90,87,87,89,88,96,84,85,82,80,93,80,84,86,81)
Gas = c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5)
Method = c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4)

#Fitting CRBD. Make sure to change variable ‘Gas’ and ‘Method’ to ‘Factor’ vari-
ables as given below:
model = aov(Obs ~ factor(Gas)+ factor(Method))
anova(model)

#Diagnostic plots
par(mfrow=c(2,2))
plot(model)

#Tukey Honestly Significant Differences
TukeyHSD(model)
```

The R output appears similar to that obtained via the MINITAB procedure. Based on this output, as we concluded earlier from the MINITAB output, we have ample evidence to reject the hypothesis that all gasoline types have equal effects. And based on a p-value analysis, we have a lack of evidence to reject the claim that the blocks have no significant effect.

17.4.1 The Friedman F_r-Test for Randomized Complete Block Design (Nonparametric Method)

The nonparametric test proposed by Friedman is an alternative to the parametric test when the normality assumption in model (17.4.1) is not met. The Friedman test is quite similar to the Kruskal–Wallis H test used for one-way experimental designs. In the Kruskal–Wallis H test, the entire N observations are ranked from 1 to N, assigning rank 1 to the smallest observation and rank N to the largest observation. In the Friedman test, observations in each block are ranked separately, assigning rank 1 to the smallest observation and rank a to the largest observation, where a is the number of treatments in each of the b blocks.

Suppose that $R_1, R_2, \ldots, R_a$ are the *rank sums of treatments.* Then the Friedman test statistic is defined as

$$F_r = \frac{12}{ba(a+1)} \sum_{i=1}^{a} \left(R_i - \frac{b(a+1)}{2} \right)^2 \tag{17.4.19}$$

The computational formula for the Friedman test statistic is given by

$$F_r = \frac{12}{ba(a+1)} \sum_{i=1}^{a} R_i^2 - 3b(a+1) \tag{17.4.20}$$

The test statistic F_r is approximately distributed as χ^2 with $(a-1)$ degrees of freedom. The hypothesis of equal treatment effects is rejected at the α level of significance if

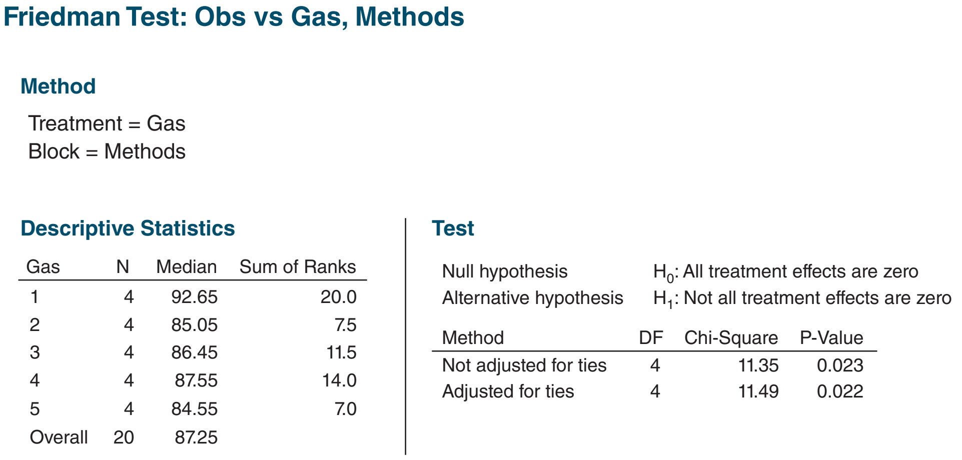
$$F_r > \chi^2_{(a-1);\alpha} \tag{17.4.21}$$

Otherwise, the hypothesis is not rejected.

Example 17.4.3 (Example 17.4.2 revisited) *Analyze the gasoline data in Table 17.4.5 using the Friedman test with MINITAB and R.*

MINITAB

1. Enter the data in column C1.
2. Enter the block identifiers (1, 2, 3, 4) in column C2.
3. Enter the treatment identifiers (1, 2, 3, 4, 5) in column C3. We name these columns as Obs., Methods, and Gas, respectively.
4. Select **Stat** > **Nonparametrics** > **Friedman . . . .**
5. In the resulting dialog box, type Obs. in the box next to **Response**, Gas in the box next to **Treatment**, and Methods in the box next to **Blocks**. Then click **OK**. The MINITAB output appears in the Session window as follows:



```
Friedman Test: Obs vs Gas, Methods

  Method

   Treatment = Gas
   Block = Methods


  Descriptive Statistics                 Test

   Gas      N    Median   Sum of Ranks     Null hypothesis          H0: All treatment effects are zero
   1        4    92.65            20.0     Alternative hypothesis   H1: Not all treatment effects are zero
   2        4    85.05             7.5
   3        4    86.45            11.5     Method                 DF   Chi-Square   P-Value
   4        4    87.55            14.0     Not adjusted for ties   4        11.35     0.023
   5        4    84.55             7.0     Adjusted for ties       4        11.49     0.022
   Overall  20   87.25
```

MINITAB, when using the Friedman test, tests only the treatment means. Our conclusion does not change even though the p-value is somewhat higher.

USING R

Solution: The R function 'friedman.test($a \sim b|c$)' can be used to conduct the required Friedman test as shown in the following R-code. Note that in the formula: $a \sim b|c$; a, b; and c are the response data, treatments, and blocks, respectively.

```
Obs = c(92,88,89,90,87,90,87,87,89,88,96,84,85,82,80,93,80,84,86,81)
Gas = c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5)
Method = c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4)

#Conduct the Friedman test
friedman.test(Obs ~ Gas|Method)

#R output
Friedman rank sum test
data: Obs and Gas and Method
Friedman chi-squared = 11.494, df = 4, p-value = 0.02154
```

17.4.2 Experiments with One Missing Observation in an RCB-Design Experiment

Consider an RCB-design experiment with a treatments and b blocks, and suppose that one observation is missing. Denote the missing observation by x, and suppose that the missing observation corresponds to the ith treatment of the jth block. Then, let $\tilde{T}_{i\cdot}, \tilde{T}_{\cdot j}$ and $\tilde{T}$ be, respectively, the ith treatment sum, the jth block sum, and sum of all nonmissing observations, we then have that

$$\tilde{T}_{i\cdot} + x = T_{i\cdot}, \quad \tilde{T}_{\cdot j} + x = T_{\cdot j}, \quad \tilde{T} + x = G$$

In other words, $\tilde{T}_{i\cdot}$ is the sum of all the $(b\ -1)$ nonmissing observations generated using the ith treatment, $\tilde{T}_{\cdot j}$ is the sum of all the $(a\ -1)$ nonmissing observations of the jth block, and $\tilde{T}$ is the sum of all the nonmissing $(ab\ -1)$ observations. Referring to Equation (17.4.18), we can verify that the terms containing x in SS_E have a value, say δ, where

$$\delta = x^2 - \frac{1}{b}(\tilde{T}_{i\cdot} + x)^2 - \frac{1}{a}(\tilde{T}_{\cdot j} + x)^2 + \frac{1}{ab}(\tilde{T} + x)^2 \tag{17.4.22}$$

Now minimizing SS_E means that we are involved in minimizing δ. Thus, differentiating δ partially with respect to x and equating to zero, we find (calling $\hat{x}$ the solution to $\partial\delta/\partial x = 0$) that

$$\hat{x} - \frac{1}{b}(\tilde{T}_{i\cdot} + \hat{x}) - \frac{1}{a}(\tilde{T}_{\cdot j} + \hat{x}) + \frac{1}{ab}(\tilde{T} + \hat{x}) = 0$$

After solving the above, we give the estimated value that replaces the missing observation as

$$\hat{x} = \frac{a\tilde{T}_{i\cdot} + b\tilde{T}_{\cdot j} - \tilde{T}}{(a-1)(b-1)} \tag{17.4.23}$$

Note that when one observation is missing, the error sum of squares degrees of freedom is reduced by 1; that is, it is equal to $(a\ -1)(b\ -1)\ -1 = ab - a - b$.

17.4.3 Experiments with Several Missing Observations in an RCB-Design Experiment

Again, we consider an RCB-design experiment, where a treatments are applied in b blocks, but with various observations missing. Let x, y, $z, \ldots$, denote the missing observations. Let $\tilde{T}_x$ denote the total of the observations present in the row that contains x, $\tilde{T}_{yz}$ denote the total of the row that contains y and z if y and z are missing in the same row, etc. Further, let $\tilde{T}'_x$, $\tilde{T}'_y$, $\ldots$, denote the totals of the columns that contain x, y, and so on. Then (see (17.4.22)), the terms containing the missing observations $x, y, z, \ldots$, in the error sum of squares are

$$\delta = x^2 + y^2 + z^2 + \cdots - \frac{1}{a}\{(\tilde{T}'_x + x)^2 + (\tilde{T}'_x + y)^2 + \cdots\}$$
$$- \frac{1}{b}\{(\tilde{T}_x + x)^2 + (\tilde{T}_{yz} + y + z)^2 + \cdots\} + \frac{1}{ab}(\tilde{T} + x + y + z + \cdots)^2$$

where $\tilde{T}$ is the grand total of all observations that are present. To minimize δ, we differentiate it partially with respect to $x, y, z, \ldots$, and equate each one of them to zero. This gives a set of least square normal equations containing $x, y, z, \ldots$. Then, the estimated values of $x, y, z, \ldots$, say $\hat{x}, \hat{y}, \hat{z}$, are obtained by solving the normal equations. We illustrate this method with Example 17.4.4 below.

Note that if r observations are missing, then the error sum of squares degrees of freedom is reduced by r, that is, $(a -1)(b -1) - r$ is now the degrees of freedom for "error."

Example 17.4.4 (Comparison of machines) *A manufacturing engineer designed an experiment to compare three new machines M_1, M_2, and M_3 with the existing machine M so that an RCB-design could be used. He used five blocks (shifts) and the characteristic measured was the total number of parts produced during a predetermined period. However, unfortunately, during that production period there were four power outages, so it was not possible to determine the parts produced during those periods. The parts produced during the rest of the periods are as shown in Table 17.4.6.*

Here, x_1, x_2, x_3, and x_4 denote the parts that would have been produced if not for the power outage. To find estimates of x_1, x_2, x_3, and x_4, we proceed, as in Section 17.4.2, where we dealt with the case of only one missing observation. We minimize SS_E and

Table 17.4.6 Coded values of parts produced (parts −400).

		Blocks					
		1	2	3	4	5	Totals
	M	19	21	17	23	15	95
Machines	M_1	22	23	x_2	21	19	$85 + x_2$
	M_2	x_1	24	26	23	18	$91 + x_1$
	M_3	28	25	x_3	x_4	20	$73 + x_3 + x_4$
Totals		$69 + x_1$	93	$43 + x_2 + x_3$	$67 + x_4$	72	$344 + x_1 + x_2 + x_3 + x_4$

note that the contributions to SS_E of the terms containing x_i have sum δ, where (see (17.4.15)–(17.4.18) and Table 17.4.6) δ is given by

$$\delta = \sum_{i=1}^{4} x_i^2 - \frac{1}{5}[(85 + x_2)^2 + (91 + x_1)^2 + (73 + x_3 + x_4)^2]$$
$$-\tfrac{1}{4}[(69 + x_1)^2 + (43 + x_2 + x_3)^2 + (67 + x_4)^2] + \tfrac{1}{20}(344 + x_1 + x_2 + x_3 + x_4)^2$$

Taking the partial derivatives $\partial\delta/\partial x_i, i = 1, 2, 3, 4$, equating to zero and denoting the solutions of the resulting equations by $\hat{x}_1, \hat{x}_2, \hat{x}_3$, and $\hat{x}_4$, we are led to the equations (after some minor algebra)

$$\begin{cases} 12\hat{x}_1 + \hat{x}_2 + \hat{x}_3 + \hat{x}_4 = 365 \\ \hat{x}_1 + 12\hat{x}_2 - 4\hat{x}_3 + \hat{x}_4 = 211 \\ \hat{x}_1 - 4\hat{x}_2 + 12\hat{x}_3 - 3\hat{x}_4 = 163 \\ \hat{x}_1 + \hat{x}_2 - 3\hat{x}_3 + 12\hat{x}_4 = 283 \end{cases} \tag{17.4.24}$$

We can solve the above equations for $\hat{X} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)'$ using matrix notation. Thus, we have

$$A\hat{X} = C, \quad A = \begin{bmatrix} 12 & 1 & 1 & 1 \\ 1 & 12 & -4 & 1 \\ 1 & -4 & 12 & -3 \\ 1 & 1 & -3 & 13 \end{bmatrix}, \quad C = [365,\ 211,\ 163,\ 283]'$$

We find that

$$\hat{X} = A^{-1}C = \begin{bmatrix} 0.0861244 & -0.0107656 & 0.013158 & -0.0095694 \\ -0.0107656 & 0.0950957 & 0.032895 & 0.0011962 \\ -0.0131579 & 0.0328947 & 0.101316 & 0.0236842 \\ -0.0095694 & 0.0011962 & 0.023684 & 0.0899522 \end{bmatrix} [365,\ 211,\ 163,\ 283]'$$
$$= \begin{bmatrix} 24.3110 \\ 21.8361 \\ 25.3553 \\ 26.0766 \end{bmatrix}$$

We note that the degrees of freedom for the error, that is, for SS_E with $\hat{X} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)'$, is now

$$(a - 1)(b - 1) - 4 = (4 - 1)(5 - 1) - 4 = 8$$

These data can be analyzed by the ordinary method of analysis of variance for RCB-designs, when $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$, $\hat{x}_4$ are used for x_1, x_2, x_3, x_4. We note again that the degrees of freedom for the error sum of squares are reduced by one degree of freedom for each missing observation, as shown above.

PRACTICE PROBLEMS FOR SECTION 17.4

1. Four different types of coatings (A, B, C, D) on memory chips are studied in three different plants (P_1, P_2, P_3). As the coatings effect might differ from plant to plant,

all different coatings are studied in all the three plants. The order of testing of the four types in each plant is completely random. The data collected are

Coating type

	A	B	C	D
P_1	5.2	4.8	5.6	4.6
P_2	4.8	5.0	4.7	5.3
P_3	4.9	5.4	4.3	5.8

(a) Analyze these data as a RCB design and state your conclusions. Use $\alpha = 0.05$.
(b) Analyze these data using Friedman's nonparametric test. Use $\alpha = 0.05$.

2. Suppose that in Problem 1, the reading on coating B in plant P_2 was not completed. Estimate the missing reading, reanalyze the data, and compare your result with that in Problem 1. Use $\alpha = 0.05$.
3. A quality engineer in a paper mill decides to test the effect of five chemicals (C_1, C_2, C_3, C_4, C_5) on coated paper produced by four machines (M_1, M_2, M_3, M_4). Since there might be variability of coated paper from machine to machine, the engineer chooses to use a RCB design, using the machines as blocks. She applies all five chemicals on paper taken from each of the four machines. Chemicals are applied in a random order. The data readings are tensile strengths:

	M_1	M_2	M_3	M_4
C_1	70	68	62	71
C_2	62	77	70	69
C_3	64	67	61	66
C_4	74	75	69	73
C_5	70	64	63	67

Analyze these data and state your conclusions. Use $\alpha = 0.10$.. Prepare residual plots and examine if the model violates any assumptions.

4. Suppose that two data points (C_3, M_2) and (C_5, M_3) in Problem 3 are missing. Estimate the missing observations and analyze the data. Use $\alpha = 0.10$.
5. Analyze the data in Problem 3 using Friedman's nonparametric test. Use $\alpha = 0.10$.
6. Analyze the data you obtained after estimating the missing observation in Problem 2 using Friedman's nonparametric test. Use $\alpha = 0.05$. Compare your result with that in Problem 2.
7. Analyze the data you obtained after estimating the missing observations in Problem 4, using Friedman's nonparametric test. Use $\alpha = 0.10$.. Compare your conclusion with that in Problem 4.

17.5 TWO-WAY EXPERIMENTAL LAYOUTS

In Sections 17.3 and 17.4, we discussed one-way experimental designs and completely randomized designs. In such experimental designs, we are interested in studying the effects of one set of treatments on a response variable of interest where treatments are just the levels of the factor under investigation. In Section 17.4, we focused on RCB designs in which we have one factor of prime interest, while the other factor is referred to blocks. In RCB designs, the blocks are used to eliminate the effects of a nuisance variable, and it may be that our only interest in studying block effects is to find out whether the creation of blocks was justified. In this section, we consider *two-way experimental layouts* also called *two-way factorial experimental designs*. Such designs come about because the experimenter wishes to study the effects on a response y of two factors, say A and B, and typically, use is made of a $(a > 1)$ levels of A and b $(b > 1)$ levels of B which are completely *crossed*; that is, all ab pairs of treatments A_iB_j are used to generate observations y_{ij} on Y (here A_i, $i = 1, \ldots, a$ denotes the a levels of A, and B_j, $j = 1, \ldots, b$ denotes the b levels of B). For example, a chemist wants to investigate the effects of a production process on the yield of a chemical and suspects that the two important factors affecting the yield are the amount of a catalyst used and the reaction time.

To investigate the effect of catalyst and reaction time on the production process, the chemist would run an experiment that varies the catalyst amount, varies the reaction time, and then observe the yield produced by these various combinations of catalyst and reaction times. We refer to the catalyst as a factor, say A, with its levels being the different amounts of catalyst used, and refer to the reaction time as factor B, with its levels being the different reaction times used in the experiment. We note that in experiments with two or more factors, the treatments are the combinations of all levels of various factors involved.

The two factors are said to be *completely crossed* if each level of one factor occurs with each level of the other factor. Further, if all the treatment combinations are replicated an equal number of times, then the data obtained are said to be *balanced*. In the present discussion, we consider experiments with balanced data only.

In a factorial experiment with balanced data, the effect of the ith level A_i of a factor A is defined as the difference between the average response when factor A is applied at the ith level, and the overall average of all responses generated by using factor A at all levels. These effects are usually referred to as the *main effects*.

It is important to recognize that sometimes the effect of the level of a factor depends on the level of another factor being applied. In other words, *the effect of the level of a factor is not the same at all levels of the other factor*. In such cases, we say that an *interaction* between the factors exists. This aspect of interaction can easily be explained graphically with the help of a simple example.

Consider an experiment with two factors, each at two levels, one at a low and the other at high setting. Suppose that we run two experiments so that the observed responses (hypothetical) are as shown in Table 17.5.1. Then, we define the interaction as follows. Effects of factor B when factor A is at low and high levels in experiment I are $58-36 = 22$ and $70-48 = 22$, respectively. The interaction between factors A and B, denoted by AB, is the average of the difference between these two effects, which in this case is $(22 - 22)/2 = 0$. Hence, in experiment I, there is no interaction between factors A and B. When there is no interaction between two factors, the two lines shown in Figure 17.5.1a are parallel. Similarly in experiment II we can see that interaction between factors A and B is $AB = ((64-30) - (60-40))/2 = 7$. That is, there is an interaction between factors

Table 17.5.1 Responses of two experiments with two factors A and B.

(a) Experiment I			(b) Experiment II		
A	B	Responses	A	B	Responses
Low	Low	36	Low	Low	40
Low	High	58	Low	High	60
High	Low	48	High	Low	30
High	High	70	High	High	64

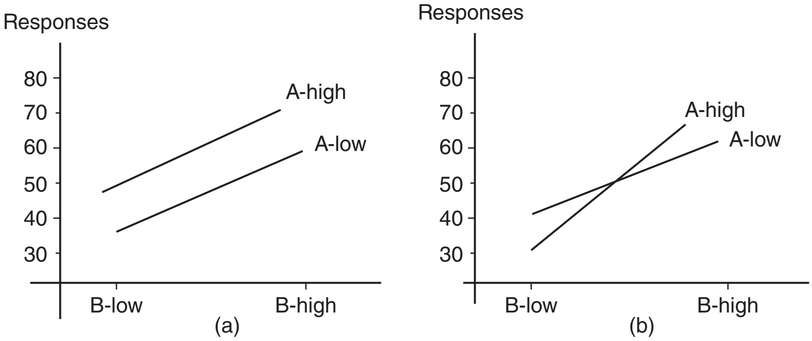


Figure 17.5.1 Plots of responses in (a) experiment I (b) experiment II.

A and B. When there is an interaction between two factors, then the two lines shown in Figure 17.5.1b, intersect. The case of interaction between two factors having three or more levels is discussed later.

The models for two-way experimental layouts with interactions and with no interactions are given in (17.5.1) and (17.5.2), respectively; with interaction, the model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, \ldots, a; \; j = 1, 2, \ldots, b; \; k = 1, 2, \ldots, r \quad (17.5.1)$$

with side conditions

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0 \qquad (17.5.1a)$$

where a and b are the number of levels of factors A and B, respectively, r is the number of replications of each treatment (i, j), and γ_{ij} is the interaction effect when factor A is applied at the ith level and factor B is applied at the jth level. Further the ε_{ijk} are assumed to be independently distributed $N(0, \sigma^2)$. If $\gamma_{ij} = 0$ for all (i, j), then the model with no interaction between A and B is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, \ 2, \ldots, a \; ; \; j = 1, \ 2, \ldots, b \; ; \; k = 1, \ 2, \ldots, r \quad (17.5.2)$$

with side conditions

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0 \qquad (17.5.2a)$$

Table 17.5.2 Observations from a two-way experimental layout with one observation per cell.

Factor B Factor A	B_1	B_2	B_3	$\cdots$	B_b	Totals	Means
A_1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1b}	$T_{1\cdot}$	$\overline{y}_{1\cdot}$
A_2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2b}	$T_{2\cdot}$	$\overline{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
A_a	y_{a1}	y_{a2}	y_{a3}	$\cdots$	y_{ab}	$T_{a\cdot}$	$\overline{y}_{a\cdot}$
Totals	$T_{\cdot1}$	$T_{\cdot2}$	$T_{\cdot3}$	$\cdots$	$T_{\cdot b}$	$T_{\cdot\cdot} = G$	
Means	$\overline{y}_{\cdot1}$	$\overline{y}_{\cdot2}$	$\overline{y}_{\cdot3}$	$\cdots$	$\overline{y}_{\cdot b}$	$\overline{y}_{\cdot\cdot}$	

where a and b are again the number of levels of factors A and B, respectively, and r is the number of replications. Further the error terms are assumed to be independently distributed as $N(0, \sigma^2)$. The model (17.5.2) is usually referred to as an *additive model* (see also after Table 17.5.2 below).

17.5.1 Two-Way Experimental Layouts with One Observation per Cell

Suppose we have two factors, A and B, at a and b levels, respectively, where factor A represents a varieties of wheat and B represents the nitrogen used at b different levels. Further, we suppose that each variety of wheat occurs exactly once with each level of nitrogen. Then, we have a completely crossed experiment with balanced data having ab observations. Each observation is specified by the levels of the two factors. The observations are usually presented in tabular form as in Table 17.5.2.

In a two-way experimental layout with one observation per cell, the model employed is given in (17.5.2), and since $r = 1$, we replace subscripts ijk on y and ε with ij. It is important to note that we are assuming that interactions play no role and are zero or negligible. If we do not assume interactions to be negligible, then there are no degrees of freedom available for the error sum of squares, so the error variance cannot be estimated. If the experimenter believes that there may be significant interactions, then the ab experiments must be replicated. However, if the experimenter can pinpoint which interactions are significant and which are not, then he or she can assume the nonsignificant interactions to be zero and use the corresponding degrees of freedom to estimate the error variance. Further, it is important to note that since the model used for two-way experimental layouts with one observation per cell is exactly the same as that used for RCB-designs, the data analysis for a two-way experimental layout with one observation per cell is carried out exactly in the same manner as for RCB-designs. However, there are two important differences between the RCB-designs and the two-way experimental layout with one observation per cell:

1. In an RCB-design, there is one factor of prime interest, while the other factor (blocks) just represents a nuisance variable. In a two-way experimental design with one observation per cell, both factors are of prime interest.
2. In an RCB-design, randomization is done within each block separately, whereas in a two-way experimental design with one observation per cell, randomization is done for the whole experiment.

17.5.2 Two-Way Experimental Layouts with $r > 1$ Observations per Cell

In a two-way experimental layout with r observations per cell, the model employed is that given in (17.5.1).

To restate, the model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots, r \tag{17.5.3}$$

with side conditions

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0 \tag{17.5.3a}$$

where a and b are the number of levels of factors A and B, respectively, r is the number of replications of all the ab treatments, and α_i, β_j, and γ_{ij} are unknown parameters. Further, the ε_{ijk} are assumed to be independently distributed as $N(0, \sigma^2)$. The observations from a two-way experimental layout are displayed in Table 17.5.3.

Table 17.5.3 Observations from a two-way experimental layout with $r > 1$ observations per cell.

Factor B Factor A	B_1	B_2	$\cdots$	B_b	Row totals	Row means
A_1	y_{111} $T_{11\cdot}$ $\vdots$ $\overline{y}_{11\cdot}$ y_{11r}	y_{121} $T_{12\cdot}$ $\vdots$ $\overline{y}_{12\cdot}$ y_{12r}	$\cdots$	y_{1b1} $T_{1b\cdot}$ $\vdots$ $\overline{y}_{1b\cdot}$ y_{1br}	$T_{1\cdot\cdot}$	$\overline{y}_{1\cdot\cdot}$
A_2	y_{211} $T_{21\cdot}$ $\vdots$ $\overline{y}_{21\cdot}$ y_{21r}	y_{221} $T_{22\cdot}$ $\vdots$ $\overline{y}_{22\cdot}$ y_{22r}	$\cdots$	y_{2b1} $T_{2b\cdot}$ $\vdots$ $\overline{y}_{2b\cdot}$ y_{2br}	$T_{2\cdot\cdot}$	$\overline{y}_{2\cdot\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
A_a	y_{a11} $T_{a1\cdot}$ $\vdots$ $\overline{y}_{a1\cdot}$ y_{a1r}	y_{a21} $T_{12\cdot}$ $\vdots$ $\overline{y}_{a2\cdot}$ y_{a2r}	$\cdots$	y_{ab1} $T_{ab\cdot}$ $\vdots$ $\overline{y}_{ab\cdot}$ y_{abr}	$T_{a\cdot\cdot}$	$\overline{y}_{a\cdot\cdot}$
Column totals	$T_{\cdot 1\cdot}$	$T_{\cdot 2\cdot}$	$\cdots$	$T_{\cdot b\cdot}$	$G = T_{\ldots}$	
Column means	$\overline{y}_{\cdot 1\cdot}$	$\overline{y}_{\cdot 2\cdot}$	$\cdots$	$\overline{y}_{\cdot b\cdot}$		$\overline{y}_{\ldots}$

$T_{ij\cdot}$ is the sum of all observations in the *ij*th cell and $\overline{y}_{ij\cdot}$ is the average of all the observations in the *ij*th cell; $T_{i\cdot\cdot}$ is the sum of all observations in the *i*th row and $\overline{y}_{i\cdot\cdot}$ is the average of all the observations in the *i*th row; $T_{\cdot j\cdot}$ is the sum of all observations in the *j*th column and $\overline{y}_{\cdot j\cdot}$ is the average of all the observations in the *j*th column; and $\overline{y}_{\ldots}$ is the grand average of all the observations with $T_{\ldots}$ (or G) denoting the (grand) sum of all the observations, so that

$$\begin{cases} \overline{y}_{ij\cdot} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk} & \text{or } \overline{y}_{ij\cdot} = \frac{1}{r}T_{ij\cdot} \quad \text{(cell averages)} \\ \overline{y}_{i\cdot\cdot} = \frac{1}{br}\sum_{j=1}^{b}\ \sum_{k=1}^{r} y_{ijk} & \text{or } \overline{y}_{i\cdot\cdot} = \frac{1}{br}T_{i\cdot\cdot} \quad \text{(row averages)} \\ \overline{y}_{\cdot j\cdot} = \frac{1}{ar}\sum_{i=1}^{a}\ \sum_{k=1}^{r} y_{ijk} & \text{or } \overline{y}_{\cdot j\cdot} = \frac{1}{ar}T_{\cdot j\cdot} \quad \text{(column averages)} \\ \overline{y}_{\ldots} = \frac{1}{abr}\sum_{i=1}^{a}\ \sum_{j=1}^{b}\ \sum_{k=1}^{r} y_{ijk} & \text{or } \overline{y}_{\ldots} = \frac{1}{abr}T_{\ldots} \quad \text{(grand averages)} \end{cases} \tag{17.5.4}$$

The error sum of squares under the model (17.5.3) is given by

$$Q = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \tag{17.5.5}$$

The least-squares estimates of μ, α_i, β_j, and γ_{ij} are obtained by minimizing the error sum of squares (17.5.5) under the model (17.5.3), subject to the conditions in (17.5.3a).

The least-squares normal equations, as in the one-way experimental layout, are obtained by equating to zero the partial derivatives of $Q = \sum_{i=1}^{a}\ \sum_{j=1}^{b}\ \sum_{k=1}^{r}$ $(Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$ with respect to μ, α_i, β_j, and γ_{ij}. Clearly, by solving these normal equations subject to the conditions in (17.5.3a), we obtain the solutions $\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j, \hat{\gamma}_{ij}$ given by

$$\hat{\mu} = \overline{y}_{\ldots}, \quad \hat{\alpha}_i = \overline{y}_{i\cdot\cdot} - \overline{y}_{\ldots}, \quad \hat{\beta}_j = \overline{y}_{\cdot j\cdot} - \overline{y}_{\ldots}, \quad \text{and} \quad \hat{\gamma}_{ij} = \overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\ldots} \tag{17.5.5a}$$

Substituting (17.5.5a) in (17.5.5), we have that Q has minimum value, say SS_E, given by

$$\text{Min Q} = SS_E = \sum_{i=1}^{a}\ \sum_{j=1}^{b}\ \sum_{k=1}^{r}(y_{ijk} - \overline{y}_{ij\cdot})^2 \tag{17.5.6}$$

Further, under the assumption that the β_j are independent and distributed as $N(0, \sigma^2)$, it can be shown that SS_E is distributed as $\sigma^2\chi^2_{ab(r-1)}$ so that

$$\hat{\sigma}^2 = S^2 = \frac{\text{Min } Q}{ab(r-1)} = \frac{SS_E}{ab(r-1)} = MS_E \tag{17.5.7}$$

is an unbiased estimate of σ^2.

The total variation sum of squares and its various components are given by

$$SS_{total} = \sum_{i=1}^{a}\ \sum_{j=1}^{b}\ \sum_{k=1}^{r}(y_{ijk} - \overline{y}_{\ldots})^2 = \sum_{i=1}^{a}\ \sum_{j=1}^{b}\ \sum_{k=1}^{r} y_{ijk}^2 - \frac{G^2}{N}, \quad N = abr \tag{17.5.8}$$

$$SS_A = br\sum_{i=1}^{a}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\ldots})^2 = br\sum_{i=1}^{a}\overline{y}_{i\cdot\cdot}^2 - \frac{G^2}{N} = \sum_{i=1}^{a}\frac{T_{i\cdot\cdot}^2}{br} - \frac{G^2}{N} \tag{17.5.9}$$

Similarly

$$SS_B = ar\sum_{j=1}^{b}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\ldots})^2 = \sum_j \frac{T_{\cdot j\cdot}^2}{ar} - \frac{G^2}{abr} \tag{17.5.10}$$

Also from (17.5.6) we can show that

$$SS_E = \sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j \frac{T_{ij\cdot}^2}{r} \tag{17.5.11}$$

Finally, we let $SS_{AB} = r\sum\sum\ \hat{\gamma}_{ij}^2$ the interaction sum of squares, so that

$$SS_{AB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\ldots})^2 \tag{17.5.11a}$$

The total sum of squares of deviations y_{ijk} from the grand mean $y_{\ldots}$ is

$$\begin{aligned} SS_{total} &= \sum_{i=1}^{a}\ \sum_{j=1}^{b}\ \sum_{k=1}^{r}(y_{ijk} - \overline{y}_{\ldots})^2 \\ &= \sum\ \sum\ \sum[(y_{ijk} - \overline{y}_{ij\cdot}) + (\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} - \overline{y}_{\ldots}) + (\overline{y}_{i\cdot\cdot} - \overline{y}_{\ldots}) + (\overline{y}_{\cdot j\cdot} - \overline{y}_{\ldots})]^2 \\ &= \sum\ \sum\ \sum(y_{ijk} - \overline{y}_{ij\cdot})^2 + \sum\ \sum\ \sum(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\ldots})^2 \\ &\quad + \sum\ \sum\ \sum(\overline{y}_{i\cdot\cdot} - \overline{y}_{\ldots})^2 + \sum\ \sum\ \sum(\overline{y}_{\cdot j\cdot} - \overline{y}_{\ldots})^2 \end{aligned}$$

Because the cross products sum to zero. Thus, we have

$$SS_{total} = SS_E + SS_{AB} + SS_A + SS_B \tag{17.5.12}$$

Thus, we have

$$SS_{AB} = SS_{total} - SS_A - SS_B - SS_E \tag{17.5.13}$$

Sometimes SS_{AB} may also be obtained using the following expression:

$$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\cdot}^2}{r} - \frac{G^2}{N} - SS_A - SS_B \tag{17.5.13a}$$

Note that SS_A, SS_B, and SS_{AB} can also be obtained as:

$$SS_A = \sum_i \hat{\alpha}_i T_{i\cdot\cdot}, \quad SS_B = \sum_j \hat{\beta}_j T_{\cdot j\cdot}, \quad SS_{AB} = \sum_i\sum_j \hat{\gamma}_{ij} T_{ij\cdot} \tag{17.5.13b}$$

where $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are the least-square estimates of α_i, β_j, and γ_{ij}, respectively, given in (17.5.5a).

Next, it can easily be shown that under normality SS_E, SS_A, SS_B, and SS_{AB} are independently distributed as $\sigma^2\chi^2_{ab(r-1)}$, $\sigma^2\chi^2_{(a-1)}(\lambda_1)$, $\sigma^2\chi^2_{(b-1)}(\lambda_2)$, and $\sigma^2\chi^2_{(a-1)(b-1)}(\lambda_3)$, respectively, where λ_1, λ_2, and λ_3 are the noncentrality parameters given by $\lambda_1 = \left(br\sum_i \alpha_i^2/\sigma^2\right)$, $\lambda_2 = \left(ar\sum_j \beta_j^2/\sigma^2\right)$, and $\lambda_3 = \left(r\sum_i\ \sum_j \gamma_{ij}^2/\sigma^2\right)$. Note, however, that under the null hypotheses to be discussed here, the noncentrality parameters are zero, so SS_A, SS_B and SS_{AB} are distributed as central Chi-squares, whereas SS_E, regardless of any hypothesis, is always distributed as a central Chi-square. Thus, we have

$$\text{if } \alpha_i = 0, \text{ all } i, \text{ then } \frac{SS_A/(a-1)}{SS_E/ab(r-1)} = \frac{MS_A}{MS_E} \sim F_{(a-1),ab(r-1)} \tag{17.5.14}$$

$$\text{if } \beta_i = 0, \text{ all } j, \text{ then } \frac{SS_B/(b-1)}{SS_E/ab(r-1)} = \frac{MS_B}{MS_E} \sim F_{(b-1),ab(r-1)} \tag{17.5.15}$$

and

$$\text{if } \alpha_{ij} = 0, \text{ all } i,j, \text{ then } \frac{SS_{AB}/(a-1)(b-1)}{SS_E/ab(r-1)} = \frac{MS_{AB}}{MS_E} \sim F_{(a-1)(b-1),ab(r-1)} \tag{17.5.16}$$

In a two-way experimental design, the hypotheses of prime interest are as follows:

1. $H_0 :\ \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ versus H_1 : not all $\alpha_i = 0$
2. $H_0 :\ \beta_1 = \beta_2 = \cdots = \beta_b = 0$ versus H_1 : not all $\beta_i = 0$
3. $H_0 :\ \gamma_{11} = \gamma_{12} = \cdots = \gamma_{ab} = 0$ versus H_1 : not all $\gamma_{ij} = 0$

Hence, we reject the hypothesis that all main effects $\alpha_i = 0$ at the α level of significance if

$$\frac{MS_A}{MS_E} \geq F_{(a-1),ab(r-1);\alpha} \tag{17.5.17}$$

Similarly, we reject the hypothesis that all main effects $\beta_j = 0$ at the α level of significance if

$$\frac{MS_B}{MS_E} \geq F_{(b-1),ab(r-1);\alpha} \tag{17.5.18}$$

Last, we reject the hypothesis that all interactions $\gamma_{ij} = 0$ at the α level of significance if

$$\frac{MS_{AB}}{MS_E} \geq F_{(a-1)(b-1),ab(r-1);\alpha}$$

These results are summarized in the ANOVA Table 17.5.4.

The two-way experimental design, $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, when the interactions are negligible, reduces to $y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$. The reduced model is usually referred to as an *additive model.*

Usually, the test of $H_0 : \gamma_{ij} = 0$ for all i, j is performed first. If the hypothesis that interactions are zero is rejected, then the main effects for each factor are tested at each level of the factor. For example, if interaction is indicated, the main effects of factor, say, A should be tested at each level of factor B. In other words, to test the main effects of factor A, we proceed as if we have one-way experimental layouts. If $H_0 : \gamma_{ij} = 0$ for all i, j is not rejected, then we proceed to test the effects of two factors using (17.5.17) and (17.5.18).

Example 17.5.1 (Operators and machines manufacturing ball-bearings) *Four operators work in turn on four machines to produce ball bearings. The data in Table 17.5.5 gives the percentage of defective ball-bearings produced by each operator using the different machines on two consecutive days. In each replication, all 16 treatments were run in random order.*

Table 17.5.4 ANOVA table for a two-way experimental design (r observations per cell).

Source	DF	SS	MS	$E(MS)$	F-ratio
A	$a-1$	SSA	$MS_A = \dfrac{SS_A}{(a-1)}$	$\sigma^2 + \dfrac{br}{a-1}\sum_{i=1}^{a}\alpha_i^2$	$\dfrac{MS_A}{MS_E}$
B	$b-1$	SSB	$MS_B = \dfrac{SS_B}{(b-1)}$	$\sigma^2 + \dfrac{ar}{b-1}\sum_{j=1}^{b}\beta_j^2$	$\dfrac{MS_B}{MS_E}$
AB	$(a-1)(b-1)$	SS_{AB}	$MS_E = \dfrac{SS_{AB}}{(a-1)(b-1)}$	$\sigma^2 + \dfrac{r}{(a-1)(b-1)}\sum_i\sum_j\gamma_{ij}^2$	$\dfrac{MS_{AB}}{MS_E}$
Error	$ab(r-1)$	SS_E	$MS_E = \dfrac{SS_E}{ab(r-1)}$	σ^2	
Total	$abr-1$	SS_{total}			

Table 17.5.5 Percentage (after coding each value by subtracting 10) of defective ball-bearings.

Operators	M_1	M_2	M_3	M_4	Row totals	Row means
O_1	4 (5) 6	2 (3) 4	0 (2) 4	2 (2) 2	$T_{1..} = 24$	$\bar{y}_{1..} = 3$
O_2	3 (4) 5	5 (4) 3	2 (2) 2	0 (2) 4	$T_{2..} = 24$	$\bar{y}_{2..} = 3$
O_3	2 (3) 4	1 (3) 5	2 (4) 6	1 (2) 3	$T_{3..} = 24$	$\bar{y}_{3..} = 3$
O_4	6 (7) 8	4 (5) 6	4 (4) 4	3 (4) 5	$T_{4..} = 40$	$\bar{y}_{4..} = 5$
Column totals	$T_{.1.} = 38$	$T_{.2.} = 30$	$T_{.3.} = 24$	$T_{.4.} = 20$	$G = T_{...} = 112$	
Column means	$\bar{y}_{.1.} = 4 \cdot 75$	$\bar{y}_{.2.} = 3 \cdot 75$	$\bar{y}_{.3.} = 3$	$\bar{y}_{.4.} = 2.5$		$\bar{y}_{...} = 3.5$

The numbers in parentheses are the cell means.

Analyze the data on defective ball-bearings in Table 17.5.5. (The numbers in parentheses are the cell means.)

Solution: Using (17.5.8)–(17.5.12), we have (A = operators, B = machines)

$$\begin{aligned}
SS_A &= \tfrac{(24)^2+(24)^2+(24)^2+(40)^2}{(4)(2)} - \tfrac{(112)^2}{(4)(4)(2)} = 24\\
SS_B &= \tfrac{(38)^2+(30)^2+(24)^2+(20)^2}{(8)} - \tfrac{(112)^2}{(32)} = 23\\
SS_E &= (4^2+6^2+\cdots+3^2+5^2) - \tfrac{1}{2}(10^2+\cdots+8^2) = 50\\
SS_{tot} &= (4^2+\cdots+3^2+5^2) - \tfrac{(112)^2}{(32)} = 110\\
SS_{AB} &= SS_{tot} - SS_A - SS_B - SS_E = 13
\end{aligned}$$

The ANOVA table is as shown in Table 17.5.6.

Since the F-ratio for the interaction is less than 1 and $F_{9,16;0.05} = 2.54$, the interactions are not significant. We now consider tests concerning main effects. The mean squares due to A as well as B can now be tested against the error mean square. Moreover, $F_{3,16;0.05} = 3.24$, which is greater than the observed value of F-ratio for A as well as for B. Therefore, the main effects due to operators and due to machines are not significant.

Table 17.5.6 ANOVA table for the data in Table 17.5.6.

Source	DF	SS	MS	F-ratio
A	3	24	8.00	2.56
B	3	23	7.67	2.45
AB	9	13	1.44	0.46
Error	16	50	3.125	
Total	31	110		

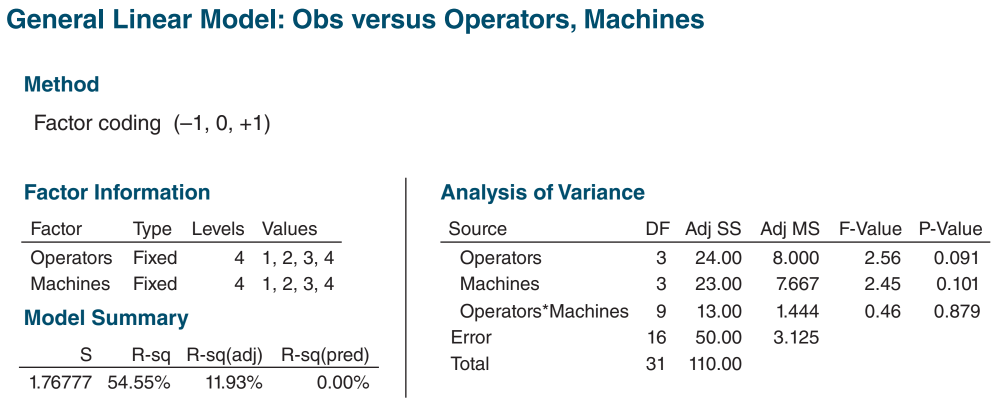
Based on the statistical procedure with significance level 0.05 we can conclude that neither operators nor machines have any significant effect on the percentage of defective ball-bearings produced.

Example 17.5.2 *Analyze the defective ball-bearings data in Example 17.5.1 using MINITAB and R.*

MINITAB

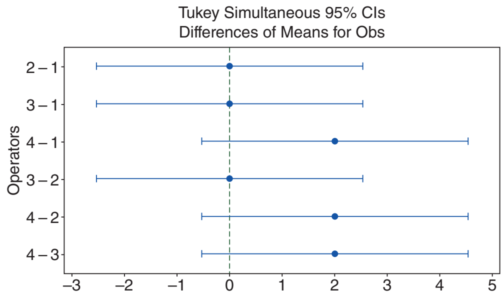
1. Enter the data in column C1.
2. Enter identifiers (1, 2, 3, 4) of treatments of factor A (Operators), in column C2.
3. Enter identifiers (1, 2, 3, 4) of treatments of factor B (Machines), in column C3.
4. We name these columns Obs., Operators, and Machines respectively.
5. Select **Stat** > **Anova** > **General Linear Model** > **Fit General Linear Model:**.
6. In the dialog box that appears type in Obs. in the box below **Response**, Operators and Machines in the box below **Factors**. Click on Model and in the dialog box that appears, add **Interactions through order: 2** by highlighting both the factors

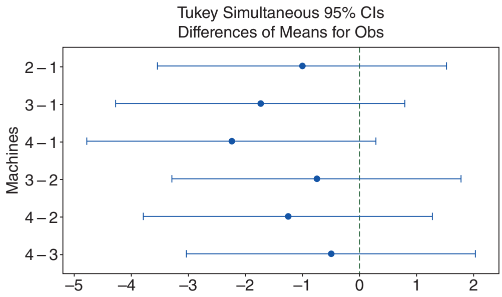
Operators and Machines. Then, click **OK**. The MINITAB output appears in the Session window as shown below:



```
General Linear Model: Obs versus Operators, Machines

Method

 Factor coding  (−1, 0, +1)

Factor Information
 Factor     Type   Levels  Values
 Operators  Fixed       4  1, 2, 3, 4
 Machines   Fixed       4  1, 2, 3, 4

Model Summary
        S     R-sq  R-sq(adj)  R-sq(pred)
 1.76777   54.55%     11.93%       0.00%

Analysis of Variance
 Source              DF  Adj SS  Adj MS  F-Value  P-Value
   Operators          3   24.00   8.000     2.56    0.091
   Machines           3   23.00   7.667     2.45    0.101
   Operators*Machines 9   13.00   1.444     0.46    0.879
 Error               16   50.00   3.125
 Total               31  110.00
```

To conduct the multiple comparisons, select **Stat** > **Anova** > **General Linear Model** > **Comparisons:**. Then select the comparison **Method** (e.g., Tukey) and **Choose terms for comparisons** (e.g., Operators) from the new window appears. Bellow is the Tukey multiple comparison MINITAB outputs for both the factors Operators and Machines.





As shown in the above Tukey multiple comparison MINITAB outputs, none of operator nor machine produces significantly different amount of defective ball-bearings since all the intervals contain zero indicating corresponding means are insignificant.

USING R

Solution: The R function 'aov()' can be used to fit the required Two-Way experimental layout as shown in the following R-code.

```
Obs = c(4,6,3,5,2,4,6,8,2,4,5,3,1,5,4,6,0,4,2,2,2,6,4,4,2,2,0,4,1,3,3,5)
Operators = c(1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4)
```

```
Machines = c(1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4)


#Fitting ANOVA model
model = aov(Obs ~ factor(Operators)*factor(Machines))
anova(model)


#Diagnostic plots
par(mfrow=c(2,2))
plot(model)


#Tukey Honestly Significant Differences
TukeyHSD(model)
```

Note that interpretation of MINITAB and R is exactly the same as given in Example 17.5.1.

Example 17.5.3 (Hydroquinine and thermometers) *Duncan (1958) quotes the following example. Three analysts, $A_1, A_2,$, and A_3, each makes two determinations of the melting point of hydroquinine (in degrees centigrade) with each of four different thermometers B_1, B_2, B_3, and B_4. Each reading minus 172° C is given in Table 17.5.7. In each replication, all 12 treatments were run in random order (the entries in parentheses are the cell means $\overline{y}_{ij}$). In this example, $a = 3$, $b = 4$, and $r = 2$.*

Solution: Using (17.5.8)–(17.5.12), we obtain

$$\begin{aligned}
SS_A &= \tfrac{9^2+1^2+8.5^2}{8} - \tfrac{(18.5)^2}{24} = 19.281 - 14.260 = 5.021\\
SS_B &= \tfrac{8^2+6^2+1^2+3.5^2}{6} - \tfrac{(18.5)^2}{24} = 18.875 - 14.260 = 4.615\\
SS_E &= 2^2 + 1.5^2 + \cdots + 0.5^2 + 1^2 - \tfrac{3.5^2+2.5^2+\cdots+1.5^2}{2} = 29.250 - 26.625 = 2.625\\
SS_{total} &= 2^2 + \cdots + 0.5^2 + 1^2 - \tfrac{18.5^2}{24} = 29.250 - 14.260 = 14.990
\end{aligned}$$

Table 17.5.7 Data on hydroquinine and thermometers.

	Thermometers				
Analyst	B_1	B_2	B_3	B_4	$T_{i\cdot\cdot}$
A_1	2.0 1.5 (1.75)	1.0 1.5 (1.25)	−0.5 0.5 (0.00)	1.5 1.5 (1.50)	9.0
A_2	1.0 1.0 (1.00)	0.0 1.0 (0.50)	−1.0 0.0 (−0.50)	−1.0 0.0 (−0.50)	1.0
A_3	1.5 1.0 (1.25)	1.0 1.5 (1.25)	1.0 1.0 (1.00)	0.5 1.0 (0.75)	8.5
$T_{\cdot j\cdot}$	8.0	6.0	1.0	3.5	18.5

Table 17.5.8 ANOVA table for the data in Table 17.5.7.

Source	DF	SS	MS	F-ratio
A	2	5.021	2.51	11.40
B	3	4.615	1.54	7.000
AB	6	2.729	0.45	2.045
Error	12	2.625	0.22	
Total	23	14.990		

Hence,

$$SS_{AB} = 14.990 - (5.021 + 4.615 + 2.625) = 14.990 - 12.261 = 2.729$$

The numerical results for the ANOVA table for this example are shown in Table 17.5.8.

We first test the hypothesis of zero interaction effects between analysts and thermometers, using the ratio of the mean squares for interaction and error. We find that the observed value of this F-ratio is

$$\frac{0.45}{0.22} = 2.045$$

Since $F_{6,12;0.05} = 2.996 \cong 3.00$, we can conclude that the interaction effects are not significantly different from zero, and we may now consider tests concerning main effects.

To test the hypothesis that effects due to analysts are zero, we use the ratio of the mean squares for the analysts and the mean square for error and find that

$$\frac{2.51}{0.22} = 11.41$$

Since $F_{2,12;0.05} = 3.885$, we reject this hypothesis to conclude that the main effects due to analysts are significantly different. To test the hypothesis of zero main effects due to thermometers, we find that the observed value of the ratio of the mean square for thermometers to the mean square for error is

$$\frac{1.54}{0.22} = 7.0$$

Since $F_{3,12;0.05} = 3.490$, we reject the hypothesis that the main effects due to thermometers are zero, and thus conclude that there are significant main effects due to thermometers. Since we have rejected the hypotheses of equal main effects, we should estimate effects and find confidence intervals for them. For example, we have

$$\hat{\alpha}_1 = \overline{Y}_{1..} - \overline{Y}_{...} = 1.125 - 0.771 = 0.354, \quad \hat{\beta}_1 = \overline{Y}_{.1.} - \overline{Y}_{...} = 1.333 - 0.771 = 0.562$$

Similarly we have

$$\hat{\alpha}_2 = -0.646, \quad \hat{\alpha}_3 = 0.2915, \quad \hat{\beta}_2 = 0.229, \quad \hat{\beta}_3 = -0.604, \quad \hat{\beta}_4 = -0.188$$

Now using MS_E as an estimator of σ^2, along with using the Student t-distribution, we can show that $100(1-\alpha)\%$ confidence intervals for α_i and β_j are given by

$$\left(\hat{\alpha}_i \pm t_{ab(r-1);\alpha/2}\sqrt{\frac{(a-1)MS_E}{abr}}\right) \tag{17.5.19}$$

$$\left(\hat{\beta}_j \pm t_{ab(r-1);\alpha/2}\sqrt{\frac{(b-1)MS_E}{abr}}\right) \tag{17.5.20}$$

It is left to the reader to compute the confidence intervals for the main effects of Example 17.5.3. Sometimes, when we do not reject the null hypothesis of interactions being zero, we may want to test a hypothesis about contrasts of the main effects. For example, suppose that B_1 and B_2 are mercury-bulb thermometers, while B_3 and B_4 are wet-bulb thermometers. Are the mercury bulbs doing the same job as the wet bulbs? To investigate this, suppose that we let $\eta = [(\beta_1+\beta_2)-(\beta_3+\beta_4)]$. That is, we choose to test a hypothesis at, say, the 0.05 level of significance

$$H_0:\ \eta = \beta_1+\beta_2-\beta_3-\beta_4 = 0 \text{ versus } H_1:\ \eta = \beta_1+\beta_2-\beta_3-\beta_4 \neq 0$$

We have that the unbiased estimator of $\eta = [(\beta_1+\beta_2)-(\beta_3+\beta_4)]$ is

$$\begin{aligned}\hat{\eta} &= [(\hat{\beta}_1+\hat{\beta}_2)-(\hat{\beta}_3+\hat{\beta}_4)] = [((\overline{Y}_{\cdot 1\cdot}-\overline{Y}_{\ldots})+(\overline{Y}_{\cdot 2\cdot}-\overline{Y}_{\ldots}))-((\overline{Y}_{\cdot 3\cdot}-\overline{Y}_{\ldots})+(\overline{Y}_{\cdot 4\cdot}-\overline{Y}_{\ldots}))]\\ &= \overline{Y}_{\cdot 1\cdot}+\overline{Y}_{\cdot 2\cdot}-\overline{Y}_{\cdot 3\cdot}-\overline{Y}_{\cdot 4\cdot}\end{aligned}$$

But observations in the jth column are independent of observations in the j' th column for all $(j,j'), j\neq j'$, so that

$$Var(\hat{\eta}) = \sum_{j=1}^{4} V(\overline{Y}_{\cdot j\cdot}) = \sum_{j=1}^{4}\frac{\sigma^2}{ar} = \frac{4}{6}\sigma^2 = \frac{2}{3}\sigma^2$$

Hence, the estimate of $Var(\hat{\eta})$ is $(2/3)\hat{\sigma}^2 = (2/3)MS_E = (2/3)(0.22) = 0.146$, and the 95% confidence interval for η is then given by

$$(\hat{\eta} \pm t_{12;0.025}\sqrt{0.146}) = (1.583 \pm 2.179\sqrt{0.146}) = (1.583 \pm 0.833) = (0.750, 2.416)$$

Since the confidence interval for η does not contain 0, we reject H_0 and conclude that there is a significant difference between mercury-bulb and wet-bulb thermometers at the 5% level of significance.

17.5.3 Blocking in Two-Way Experimental Layouts

This section is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

17.5.4 Extending Two-Way Experimental Designs to n-Way Experimental Layouts

This section is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

PRACTICE PROBLEMS FOR SECTION 17.5

1. A manufacturing engineer uses four different raw materials (R_1, R_2, R_3, R_4) and three different temperatures (T_1, T_2, T_3) to produce copper wires used in electric cables. The engineer wishes to study the effect of different raw materials and three different temperatures on the tensile strength of the wire. She replicates the whole experiment twice and obtains the data given below. Give the mathematical model for the study in this problem and prepare the ANOVA table.

	R_1	R_2	R_3	R_4
T_1	79	72	69	73
	76	73	78	80
T_2	68	74	77	71
	72	78	72	69
T_3	75	82	76	82
	71	79	70	79

2. Refer to the data in Problem 1. Analyze these data using $\alpha = 0.05$. Is there a significant interaction effect? State the hypothesis that leads to a test that the effects of raw materials are the same at the 5% level of significance and perform the test. Give your conclusions.
3. In Problem 1, test the hypothesis that the effects of temperature are the same at the 5% level of significance. If any of the hypotheses in this problem and in Problem 2 are rejected, then estimate the corresponding effects and find 95% confidence intervals for these effects.
4. Suppose that the effects of the raw materials R_j in Problem 1 are denoted by $\beta_j, j = 1, 2, 3, 4$. Consider the following orthogonal contrasts on the data in that problem:

$$\xi_1 = \beta_1 - \beta_2, \quad \xi_2 = \beta_1 + \beta_2 - 2\beta_3, \quad \xi_3 = \beta_1 + \beta_2 + \beta_3 - 3\beta_4$$

   Use the mean square error term of the ANOVA table in Problem 1 to separately test the hypotheses $\xi_i = 0$, $i = 1, 2, 3$, each at the 5% level of significance.
5. Refer to Problem 1. Suppose that a manufacturing engineer conducted an experiment on copper wire in two different plants so that the experiment was replicated using the two plants. Since there might be some variability from plant to plant, she regarded each plant as a block. Write down the data in Problem 1 in two blocks and then reanalyze these data. Regard the upper figure in the $(T_i,\ R_j)$ cell as an observation belonging to plant I (replication 1) and the lower figure as an observation belonging to plant II (replication 2). Use $\alpha = 0.05$.
6. An experiment was performed to study the effect of plate temperature and filament lighting on transconductance of a certain type of tube. Two levels of plate

temperature (550 and 600 °F) and four levels of filament lighting current L_1, L_2, L_3, and L_4 were used; three replicates were made for each combination of plate temperature and filament current. The transconductance measurements are given below (from Bowker and Lieberman, 1959). Give an appropriate mathematical model for this study and perform a complete analysis of variance of these data. Find the p-value for the F-statistics you used for testing each of the usual hypotheses. (To compute the exact p-value, you will need to use one of the statistical packages.)

	Filament lighting current			
Plate temperature	L_1	L_2	L_3	L_4
T_1 (550°F)	3774 4364 4374	4710 4180 4514	4176 4140 4398	4540 4530 3964
T_2 (600°F)	4216 4524 4136	3828 4170 4180	4122 4280 4226	4484 4332 4390

7. Suppose that the three replications of the experiment in Problem 6 were carried out on three different days. Since there might be some variability from day to day, the experimenter regarded each day as a block. Write down the data in Problem 6 in three blocks and then reanalyze these data. Regard the upper figure in the (T_i, L_j) cell as an observation belonging to day I (replication 1), the middle figure as to day II (replication 2), and the lower figure as belonging to day III (replication 3).
8. An experiment was run to determine the effect of three types of oil, *a, b, c* on the wear of four kinds of piston rings, *A, B, C, D*. The measure of wear was taken as the logarithm of loss in piston-ring weight (in grams times 100) in a 12-hours test run. The results of the experiment are shown below:

	Oil type		
Piston ring type	*a*	*b*	*c*
A	1.782	1.568	1.570
B	1.306	1.223	1.240
C	1.149	1.029	1.068
D	1.025	1.919	1.982

Prepare the ANOVA table for these data and perform the usual testing of hypotheses. Using the S-method and T-method, find 95% confidence intervals for the contrast $\beta_1 - 2\beta_2 + \beta_3$, where β_1, β_2, and β_3 are the effects of oil types. Compare the size of the confidence intervals obtained using the S-method and the T-method.

9. Measurement of "filling time" in minutes for specimens of cloths A, B, and C taken from machines 1, 2, . . . , 9 in a certain plant gave the results shown below (actual measurement −78.00).

	Machine								
Cloth	1	2	3	4	5	6	7	8	9
A	18.76	20.69	19.77	19.85	22.28	20.39	24.31	22.90	19.28
	21.18	23.20	23.94	18.92	20.45	21.80	26.29	25.42	22.04
B	21.18	16.85	20.75	18.72	18.97	19.52	20.08	21.27	15.67
	19.10	20.16	21.49	16.14	20.31	21.27	19.36	17.82	18.84
C	21.74	22.68	21.90	20.28	19.89	21.12	23.02	27.48	18.70
	18.99	23.59	18.61	18.71	18.36	18.59	18.85	22.95	23.39

(a) Prepare the ANOVA table for this set of data.

(b) Test the hypothesis that all interactions are zero. Use $\alpha = 0.05$.

(c) If the hypothesis in (b) is not rejected, then test the hypotheses that machine effects and cloth effects are zero. Find the p-value for the F-statistics you used for testing each of these hypotheses.

10. The quality control department of a fabric-finishing plant is studying the effect of several factors on the dyeing of cotton synthetic cloth used in men's shirts. Three operators, three cycle times, and two temperatures were selected, and three small specimens of cloth were dyed under each set of conditions. The finished cloth was compared to a standard and a numerical score was assigned. The results are given in the table below. Analyze the data and state your conclusions. Comment on the model's adequacy (data from Montgomery, 2009a,b, used with permission).

	Operators at 300°C			Operators at 350°C		
Cycle time	1	2	3	1	2	3
40	23	27	31	24	38	34
	24	28	32	23	36	36
	25	26	29	28	35	39
50	36	34	33	37	34	34
	35	38	34	39	38	36
	36	39	35	35	36	31
60	28	35	26	26	36	28
	24	35	27	29	37	26
	27	34	25	35	34	24

17.6 LATIN SQUARE DESIGNS

In Section 17.4, we discussed RCB designs where a division was made on all the experimental units into different blocks, such that the experimental units were homogeneous, but there could be heterogeneity between blocks. By doing so, we were able to eliminate the effects of one nuisance variable. However, quite often there is a need to eliminate the effects of two nuisance variables. For example, in a manufacturing process the difference between machines and operators may be considered as two nuisance variables, and in agricultural experiments we may have unwanted effects due to the fertility of the plots that usually run in two directions; in pharmaceutical experiments, patients are being treated with certain drugs of different doses in different hospitals, so the difference between the patients and the hospitals are two nuisance variables.

In each case, we want to eliminate the effects of the two nuisance variables. A fairly simple experimental design for such a situation is the Latin square design. A Latin square design consists of RCBs in two directions, since each row and each column is a RCB. In a Latin square design, two nuisance variables or factors, say A and B, should be identified as *blocking* variables, or, sources of unwanted variability that cannot be excluded from the experimental environment. The factor C denotes the *studied* variable whose influence on the response is of primary concern to the experimenter. We say that a $r \times r$ Latin square design is an arrangement of r Latin letters in a $r \times r$ square array such that each Latin letter occurs once in each row and once in each column. For example, a 4×4 Latin square is shown here:

A	*B*	*C*	*D*
B	*C*	*D*	*A*
C	*D*	*A*	*B*
D	*A*	*B*	*C*

A Latin square is called a *standard Latin square* if the letters in the first row and the first column are in alphabetical order. For example, the only possible 4×4 standard Latin squares are

A	*B*	*C*	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
B	*C*	*D*	*A*	*B*	*A*	*D*	*C*	*B*	*D*	*A*	*C*	*B*	*A*	*D*	*C*
C	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*A*	*D*	*B*	*C*	*D*	*B*	*A*
D	*A*	*B*	*C*	*D*	*C*	*B*	*A*	*D*	*C*	*B*	*A*	*D*	*C*	*A*	*B*

The total number of 4×4 Latin squares is 576. These Latin squares can be obtained from the standard Latin squares by first permuting all (entire) four columns in 4! ways and then, in each of these, permuting the last three rows in 3! ways, so that one standard Latin square yields $4! \times 3! = 144$ different Latin squares. Thus, four standard Latin squares yield a total of $4 \times 144 = 576$ different Latin squares.

An advantage of a $r \times r$ Latin square design is that it allows us to study three factors (each factor at r levels) using only r^2 treatments. That is, this design uses only $(1/r)$ of the treatments that would be needed to run an experiment using a completely crossed three-way experimental design. However, the disadvantages of Latin square designs are (i) the model for a Latin square design is additive—that is, *all interactions are assumed to be zero*—and (ii) the Latin square puts a restriction on each factor such that each must have r levels. This sometimes can cause problems. For example, we may not have enough resources or raw materials, say, to run the experiment at all levels of a certain factor.

Let y_{ijk} be the observed value corresponding to the ith level of factor A, jth level of factor B, and kth level of factor C. Since the kth level of factor C may or may not be present in the ij cell, the treatment (i, j, k) is present only in one cell. Thus, the model for a Latin square design is given by

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}, \quad (i, j, k) \in \Omega \tag{17.6.1}$$

where Ω is a set of r^2 values assumed by (i, j, k). This is illustrated in the 3×3 Latin square below. The side conditions for model (17.6.1) are

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0 \tag{17.6.1a}$$

and we also assume that ε_{ijk} are independent $N(0, \sigma^2)$.

Under the model (17.6.1) and (17.6.1a) the estimates of μ, α_i, β_j, and γ_k are obtained by minimizing the error sum of squares Q given by

$$Q = \sum_{(i,j,k)\in\Omega} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_k)^2 \tag{17.6.2}$$

subject to conditions (17.6.1a). On equating to zero the partial derivatives of Q with respect to μ, α_i, β_j, and γ_k, we obtain the least-square normal equations. By solving these normal equations, we obtain the estimators of μ, α_i, β_j, and γ_k. That is,

$$\hat{\mu} = \bar{y}_{\ldots}, \quad \hat{\alpha}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\ldots}, \quad \hat{\beta}_j = \bar{y}_{\cdot j \cdot} - \bar{y}_{\ldots}, \text{ and } \quad \hat{\gamma}_k = \bar{y}_{\cdot\cdot k} - \bar{y}_{\ldots} \tag{17.6.3}$$

where

$$\bar{y}_{\ldots} = \frac{1}{r^2} \sum_{(i,j,k)\in\Omega} y_{ijk}; \;\; y_{i\cdot\cdot} = \frac{1}{r} \sum_j \sum_k y_{ijk}, (i, j, k) \in \Omega; \;\; \bar{y}_{\cdot j \cdot} = \frac{1}{r} \sum_j \sum_k y_{ijk}, (i, j, k) \in \Omega$$

and

$$\bar{y}_{\cdot\cdot k} = \frac{1}{r} \sum_i \sum_j y_{ijk}, (i, j, k) \in \Omega$$

We illustrate some of these operations with the help of a 3×3 Latin square. Suppose that we assign the levels of factor A and B to the rows and columns, and the levels of C to the induced cells, so that the 3×3 Latin square design would appear as

	B_1	B_2	B_3
A_1	C_1	C_2	C_3
A_2	C_2	C_3	C_1
A_3	C_3	C_1	C_2

The $3^2 = 9$ treatments used in this design are (reading across rows)

$$\{A_1B_1C_1, \; A_1B_2C_2, \; A_1B_3C_3, \; A_2B_1C_2, \; A_2B_2C_3, \; A_2B_3C_1, \; A_3B_1C_3, \; A_3B_2C_1, \; A_3B_3C_2\}$$

so the observations obtained are

$$\{y_{111}, \; y_{122}, \; y_{133}, \; y_{212}, \; y_{223}, \; y_{231}, \; y_{313}, \; y_{321}, \; y_{332}\}$$

The index set Ω of nine values $(3 \times 3 = 9)$ assumed by (i, j, k) is

$$\Omega = \{(1,1,1), (1,2,2), (1,3,3), (2,1,2), (2,2,3), (2,3,1), (3,1,3), (3,2,1), (3,3,2)\}$$

For example, we have for the sum of all observations generated by A_1 that

$$T_{1\cdot\cdot} = y_{111} + y_{122} + y_{133}, \quad \overline{y}_{1\cdot\cdot} = \frac{1}{3}(y_{111} + y_{122} + y_{133}) = \frac{1}{3}T_{1\cdot\cdot}$$

Let $\sum\ \sum\ \sum$ denote $\sum\sum\sum_{(i,j,k)\in\Omega}$, where Ω is the set of r^2 values indexing the treatments used.

We also define various totals as follows:

$T_{i\cdot\cdot}$ = Total of ith row; $\overline{y}_{i\cdot\cdot} = T_{i\cdot\cdot}/r$, mean of the ith row
$T_{\cdot j\cdot}$ = Total of jth column; $\overline{y}_{\cdot j\cdot} = T_{\cdot j\cdot}/r$, mean of the jth column
$T_{\cdot\cdot k}$ = Total of cells containing the treatment C_k; $\overline{y}_{\cdot\cdot k} = T_{\cdot\cdot k}/r$, mean of the treatment C_k observations
$T_{\cdots}$ = Total of all r^2 observation values; $\overline{y}_{\cdots} = T_{\cdots}/r^2 = G/r^2$, mean of all r^2 observations.

Then for the $r \times r$ Latin square, the total variation sum of squares and the variation sum of squares due to rows, columns, and treatments are given by

$$SS_{total} = \sum\ \sum\ \sum (y_{ijk} - \overline{y}_{\cdots})^2 = \sum\ \sum\ \sum y_{ijk}^2 - \frac{T_{\cdots}^2}{r^2}, \quad (i,j,k) \in \Omega \tag{17.6.4}$$

$$SS_A = \sum_{i=1}^{r} (\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2 = \frac{1}{r}\sum_{i=1}^{r} T_{i\cdot\cdot}^2 - \frac{T_{\cdots}^2}{r^2} \tag{17.6.5}$$

$$SS_B = \sum_{j=1}^{r} (\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2 = \frac{1}{r}\sum_{j=1}^{r} T_{\cdot j\cdot}^2 - \frac{T_{\cdots}^2}{r^2} \tag{17.6.6}$$

$$SS_C = \sum_{k=1}^{r} (\overline{y}_{\cdot\cdot k} - \overline{y}_{\cdots})^2 = \frac{1}{r}\sum_{k=1}^{r} T_{\cdot\cdot k}^2 - \frac{T_{\cdots}^2}{r^2} \tag{17.6.7}$$

From Equation (17.6.2) we can easily see that Min$Q$ is given by

$$SS_E = \text{Min}Q = \sum\sum\sum(y_{ijk} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} - \overline{y}_{\cdot\cdot k} + 2\overline{y}_{\cdots})^2 = SS_{total} - SS_A - SS_B - SS_C \tag{17.6.8}$$

Further, under the normality assumption, SS_E and each of SS_A, SS_B, and SS_C are independently distributed as $\sigma^2\chi^2_{(r-1)(r-2)}$, $\sigma^2\chi^2_{r-1}$, respectively, when the hypotheses stated below in (17.6.10)–(17.6.12) hold. The preceding results are summarized in Table 17.6.1.

The estimator of the error variance σ^2 is given by

$$\hat{\sigma}^2 = S^2 = MS_E = SS_E/(r-1)(r-2) \tag{17.6.9}$$

Usually, the hypothesis of prime interest is (see (17.6.1))

$$H_0 : \ \gamma_1 = \gamma_2 = \cdots = \gamma_r = 0 \text{ versus } H_1 : \text{ not all } \gamma \text{ are zero} \tag{17.6.10}$$

Other possible hypotheses of interest are

$$H_0 : \ \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \text{ versus } H_1 : \text{ not all } \alpha \text{ are zero} \tag{17.6.11}$$

Table 17.6.1 ANOVA table for an $r \times r$ Latin square design.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-ratio tests
Due to factor A	SS_A	$r-1$	$MS_A = SS_A/(r-1)$	SS_A/MS_E
Due to factor B	SS_B	$r-1$	$MS_B = SS_B/(r-1)$	SS_B/MS_E
Due to factor C	SS_C	$r-1$	$MS_C = SS_C/(r-1)$	SS_C/MS_E
Error	SS_E	$(r-1)(r-2)$	$MS_E = SS_E/(r-1)(r-2)$	
Total	SS_{total}	r^2-1		

and

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_r = 0 \text{ versus } H_1: \text{ not all } \beta \text{ are zero} \tag{17.6.12}$$

If any of the hypotheses in (17.6.10)–(17.6.12) are rejected, the estimators of the corresponding parameters are (as given in (17.6.3)) computed.

Example 17.6.1 (Radioactive counting rate) *A radioactive counting rate experiment was performed on four specimens of radium C_1, C_2, C_3, and C_4. The four different specimens were subjected to a counter with four shielding methods B_1, B_2, B_3, and B_4 in various orders A_1, A_2, A_3, and A_4, so that the entire experiment was conducted according to a Latin square design. The Latin square design adopted and the observations obtained, together with certain calculations, are shown in Table 17.6.2. The variable of most interest is the radium specimen. All 16 treatments were run in random order.*

Solution: Using Equations (17.6.4)–(17.6.8), we find that various sums of squares are given by

$$\begin{aligned}
SS_A &= \tfrac{(113.04)^2}{4} + \tfrac{(112.71)^2}{4} + \tfrac{(113.75)^2}{4} + \tfrac{(113.22)^2}{4} - \tfrac{(452.72)^2}{16} = 0.14 \\
SS_B &= \tfrac{(112.73)^2}{4} + \tfrac{(114.16)^2}{4} + \tfrac{(113.14)^2}{4} + \tfrac{(112.69)^2}{4} - \tfrac{(452.72)^2}{16} = 0.35 \\
SS_C &= \tfrac{(106.45)^2}{4} + \tfrac{(111.33)^2}{4} + \tfrac{(117.81)^2}{4} + \tfrac{(117.13)^2}{4} - \tfrac{(452.72)^2}{16} = 21.44 \\
SS_{total} &= 12,831.6948 - \tfrac{(452.72)^2}{16} = 21.98 \\
SS_E &= 21.98 - (0.14 + 0.35 + 21.44) = 0.05
\end{aligned}$$

We summarize the numerical results in the ANOVA Table 17.6.3. The calculation of *SS* is aided by computations appearing in the latter part of Table 17.6.2.

Note that the 5% and 1% significance points of $F_{3,6}$ are 4.757 and 9.780, respectively. Hence, order effects are barely significant at the 5% level and not significant at the 1% level, but the shielding method effects and specimen effects are both highly significant. It is left to the reader to compute the estimates of the various parameters in the model including the error variance.

PRACTICE PROBLEMS FOR SECTION 17.6

1. A chemist is interested in studying the effects of four catalysts (C_1, C_2, C_3, C_4), four temperatures (T_1, T_2, T_3, T_4), and four allowed reaction times (R_1, R_2, R_3, R_4) on the production of a chemical. Since each experiment is very expensive and

Table 17.6.2 Radioactive counting rate data.

	Shielding method				
Order	B_1	B_2	B_3	B_4	Row totals
A_1	(C_1) 26.46	(C_3) 29.61	(C_2) 27.82	(C_4) 29.15	$T_{1..} = 113.04$
A_2	(C_2) 27.58	(C_4) 29.52	(C_1) 26.48	(C_3) 29.13	$T_{2..} = 112.71$
A_3	(C_3) 29.54	(C_1) 27.00	(C_4) 29.31	(C_2) 27.90	$T_{3..} = 113.75$
A_4	(C_4) 29.15	(C_2) 28.03	(C_3) 29.53	(C_1) 26.51	$T_{4..} = 113.22$
Totals	$T_{.1.} = 112 \cdot 73$	$T_{.2.} = 114.16$	$T_{.3.} = 113.14$	$T_{.4.} = 112.69$	$T_{...} = 452.72$

Specimen or treatment totals are summarized as follows

	C_1	C_2	C_3	C_4
	26.46	27.58	29.54	29.15
	27.00	28.03	29.61	29.52
	26.48	27.82	29.53	29.31
	26.51	27.90	29.13	29.15
Totals	$T_{..1} = 106.45$	$T_{..2} = 111.33$	$T_{..3} = 117.81$	$T_{..4} = 117.13$

Table 17.6.3 ANOVA for the data in Example 17.6.1.

Source	DF	SS	MS	F-ratio
Orders	3	0.14	0.047	5.66
Shielding methods	3	0.35	0.117	14.096
Specimens	3	21.44	7.147	861.08
Error	6	0.05	0.0083	
Total	15	21.98		

time-consuming, the chemist decides to run the experiment as a Latin square, so the whole experiment is completed using a small number of experiments. The data collected are given below. Prepare the ANOVA table for these data.

	T_1	T_2	T_3	T_4
C_1	$R_1(11)$	$R_2(19)$	$R_3(17)$	$R_4(17)$
C_2	$R_4(11)$	$R_1(13)$	$R_2(12)$	$R_3(19)$
C_3	$R_3(11)$	$R_4(11)$	$R_1(14)$	$R_2(19)$
C_4	$R_2(14)$	$R_3(18)$	$R_4(13)$	$R_1(15)$

2. Give the mathematical model you used in Problem 1. State the usual hypotheses clearly in terms of the model parameter, perform the tests, and state your conclusions. Use $\alpha = 0.05$.
3. A quality control engineer in a paper company wishes to investigate the effects of three different machines, three operators, and three amounts of bleach (used in the pulp) on the tearing strength of paper produced by the company. To carry out this experiment, the engineer decided to use a Latin square design. The data obtained from this experiment are given below; the numbers in parentheses are the levels of bleach. Prepare the ANOVA table for these data and test all appropriate hypotheses. Use $\alpha = 0.05$.

		Operators		
		O_1	O_2	O_3
	M_1	14(1)	16(2)	15(3)
Machines	M_2	18(2)	20(3)	17(1)
	M_3	16(3)	19(1)	20(2)

4. Measurements on the length in centimeters of a part from five vendors (A, B, C, D, E) are made by five different technicians (I, II, III, IV, V) who use five different instruments (1, 2, 3, 4, 5). The experiment is run as a Latin square so that all three factors are appropriately controlled. The data collected is given below.

	I	II	III	IV	V
1	A 15.1	B 15.8	C 15.2	D 16.1	E 15.8
2	B 16.2	C 15.8	D 16.3	E 15.1	A 16.0
3	C 15.6	D 15.8	E 15.9	A 15.3	B 15.5
4	D 15.9	E 16.0	A 15.5	B 16.2	C 15.4
5	E 15.6	A 15.5	B 15.7	C 15.6	D 15.8

(a) Prepare the ANOVA table for these data and perform separate tests of hypotheses for equal vendor effects, technician effects, and instrument effects. Use $\alpha = 0.05$.

(b) If any of the hypotheses in (a) is rejected, then estimate the corresponding effects.

17.7 RANDOM-EFFECTS AND MIXED-EFFECTS MODELS

17.7.1 Random-Effects Model

So far in this chapter, we have studied design models where the parameters α_i's, β_j's, ..., were all unknown constants. In this section, we consider the case where these parameters are random variables. Such a model is known as a *random-effects model.* We consider here only the case of the two-way experimental design, but this model can easily be extended for higher-order experimental designs. We consider the following model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, \ldots, a; \; j = 1, 2, \ldots, b; \; k = 1, 2, \ldots, r \quad (17.7.1)$$

where we assume that $\alpha_i, \beta_j, \gamma_{ij}$, and ε_{ijk} are independently normally distributed with mean zero and variances $\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma$, and σ^2_ε, respectively. The variance of Y_{ijk} is thus

$$\sigma^2_y = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\gamma + \sigma^2_\varepsilon \quad (17.7.2)$$

The quantities $\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma$, and σ^2_ε are usually known as *variance components.* The underlying assumptions about μ and the random errors ε_{ijk} are the same in both the fixed-effects and random-effects models. The random-effects model (17.7.1) differs from the corresponding fixed-effects model (17.5.1) in the following way. The treatments in model (17.5.1) are fixed and any conclusions made about the effects of such treatments are applicable only for the *selected treatments*, whereas in a random-effects model, treatments are randomly selected from a large population of treatments. Moreover, in the random-effects model, any conclusions made about the selected treatments are extended to the rest of the treatments in that population.

For example, suppose that the effects of both machines and operators on a production process in a manufacturing company are of interest. Suppose that we select a random sample of operators from the large population of operators and a random sample of machines from large set of machines available in the company. Then, any conclusions made about selected operators and machines will be extended to all operators and all the machines in the company. Since in a random-effects model the means of all treatments effects are assumed to be zero, any mean effect due to treatments is incorporated in the general mean μ. In the fixed-effects model, the mean of the Y_{ijk} varies, whereas in a random-effects model, the mean of all the Y_{ijk}'s is equal to μ. Further, in the random-effects model, the parameters of interest are the variance components $\sigma^2_\alpha, \sigma^2_\beta$, and σ^2_γ. Thus, for example, $\sigma^2_\alpha = 0$ implies that the effects due to treatments of a factor are the same. In the fixed-effects model, all hypotheses about treatment effects are tested against SS_E, but as we will see, in the random-effects model, testing varies; that is, not all hypotheses about treatment effects are tested against SS_E. Let

$$\overline{\alpha} = \frac{1}{a}\sum_i \alpha_i, \quad \overline{\beta} = \frac{1}{b}\sum_j \beta_j, \quad \overline{\gamma}_{i\cdot} = \frac{1}{b}\sum_j \gamma_{ij}, \quad \overline{\gamma}_{\cdot j} = \frac{1}{a}\sum_j \gamma_{ij}$$

Then from (17.7.1) we have, using the usual notation, that

$$\begin{aligned} E(SS_A) &= brE\left(\sum_i (\overline{y}_{i..} - \overline{y}_{...})^2\right) \\ &= brE\left(\sum_i (\mu + \alpha_i + \overline{\beta}_{.} + \overline{\gamma}_{i\cdot} + \overline{\varepsilon}_{i..} - \mu - \overline{\alpha}_{.} - \overline{\beta}_{.} - \overline{\gamma}_{..} - \overline{\varepsilon}_{...})^2\right) \\ &= brE\left(\sum_i (\alpha_i - \overline{\alpha}_{.})^2 + \sum_i (\overline{\gamma}_{i\cdot} - \overline{\gamma}_{..})^2 + \sum_i (\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...})^2\right) \end{aligned} \qquad (17.7.3)$$

since the expected value of cross-product terms vanishes. Further, by applying the normality assumptions, we can easily verify that

$$\sum_i (\alpha_i - \overline{\alpha}_{.})^2 \sim \sigma_\alpha^2 \chi_{a-1}^2, \quad \sum_i (\overline{\gamma}_{i\cdot} - \overline{\gamma}_{..})^2 \sim \frac{\sigma_\gamma^2}{b}\chi_{a-1}^2, \quad \text{and} \quad \sum_i (\overline{\varepsilon}_{i..} - \overline{\varepsilon}_{...})^2 \sim \frac{\sigma_\varepsilon^2}{br}\chi_{a-1}^2 \qquad (17.7.4)$$

Combining the results in (17.7.3) and (17.7.4), we have

$$E(SS_A) = br\left\{[(a-1)\sigma_\alpha^2 + (a-1)\frac{\sigma_\gamma^2}{b} + \frac{(a-1)}{br}\sigma_\varepsilon^2\right\} = (a-1)\{\sigma_\varepsilon^2 + br\sigma_\alpha^2 + r\sigma_\gamma^2\} \qquad (17.7.5)$$

Similarly, we can show that

$$\begin{aligned} E(SS_B) &= (b-1)\{\sigma_\varepsilon^2 + ar\sigma_\beta^2 + r\sigma_\gamma^2\} \\ E(SS_{AB}) &= (a-1)(b-1)\{\sigma_\varepsilon^2 + r\sigma_\gamma^2\} \end{aligned}$$

These results are summarized in Table 17.7.1.

Table 17.7.1 ANOVA for a two-way experimental design with random effects.

Source	DF	SS	MS	$E(MS)$
A	$(a-1)$	$br\sum(\overline{y}_{i..} - \overline{y})^2$	$MS_A = SS_A/(a-1)$	$\sigma_\varepsilon^2 + br\sigma_\alpha^2 + r\sigma_\gamma^2$
B	$(b-1)$	$ar\sum(\overline{y}_{.j\cdot} - \overline{y})^2$	$MS_B = SS_B/(b-1)$	$\sigma_\varepsilon^2 + ar\sigma_\beta^2 + r\sigma_\gamma^2$
AB	$(a-1)(b-1)$	By subtraction	$MS_{AB} = SS_{AB}/(a-1)(b-1)$	$\sigma_\varepsilon^2 + r\sigma_\gamma^2$
Error	$ab(r-1)$	$\sum\sum\sum(y_{ijk} - \overline{y}_{ij\cdot})^2$	$MS_E = SS_E/ab(r-1)$	σ_ε^2
Total	$abr-1$	$\sum y_{ijk}^2 - abr\overline{y}_{...}^2$		

The hypotheses of general interest are

$$H_0 : \ \sigma_\gamma = 0 \,\text{versus}\, H_1 : \ \sigma_\gamma > 0 \qquad (17.7.6)$$

$$H_0 : \ \sigma_\alpha = 0 \,\text{versus}\, H_1 : \ \sigma_\alpha > 0 \qquad (17.7.6a)$$

$$H_0 : \ \sigma_\beta = 0 \,\text{versus}\, H_1 : \ \sigma_\beta > 0 \qquad (17.7.6b)$$

Note that under the null hypothesis H_0 of (17.7.6),

$$\frac{MS_{AB}}{MS_E} \sim F_{(a-1)(b-1),ab(r-1)}$$

Thus, we reject the null hypothesis H_0 of (17.7.6) at the α level of significance if

$$\frac{MS_{AB}}{MS_E} > F_{(a-1)(b-1),ab(r-1);\alpha}$$

If H_0 of (17.7.6) is *rejected*, then all the interactions γ_{ij} are not the same, and the statistics used for testing (17.7.6a) and (17.7.6b) are, respectively, MS_A/MS_{AB} and MS_B/MS_{AB} (see the E(MS) column of Table 17.7.1). If (17.7.6) is *not rejected*, then the statistics used for testing (17.7.6a) and (17.7.6b) are, respectively, MS_A/MS_E and MS_B/MS_E.

Often, an experimenter is most interested in estimating the various variance components and determining the percentage of variation of the total variation contributed by each. We can easily see from the ANOVA table (Table 17.7.1) that the estimators of various components are given by

$$\hat{\sigma}^2_\varepsilon = MS_E \tag{17.7.7}$$

$$\hat{\sigma}^2_\gamma = \text{Max}\left\{0, \frac{MS_{AB} - MS_E}{r}\right\} \tag{17.7.7a}$$

$$\hat{\sigma}^2_\beta = \text{Max}\left\{0, \frac{MS_B - MS_{AB}}{ar}\right\} \tag{17.7.7b}$$

$$\hat{\sigma}^2_\alpha = \text{Max}\left\{0, \frac{MS_A - MS_{AB}}{br}\right\} \tag{17.7.7c}$$

17.7.2 Mixed-Effects Model

We now consider the case when some of the parameters in a model are fixed unknown constants and the rest are random variables. Such a model is called a *mixed-effects model.* We consider an experiment with two factors, say A and B, and assume that A has fixed treatment effects, while B is a random sample of b treatments from a large set of treatments. The model used for this type of experiment is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, \ldots, a; \quad j = 1, 2, \ldots, b; \quad k = 1, 2, \ldots, r \tag{17.7.8}$$

where α_i are unknown constants satisfying the side condition $\sum \alpha_i = 0$, and the β_j, γ_{ij} (for fixed j) and ε_{ijk} are independently and normally distributed with mean zero and variances σ^2_β, σ^2_γ, and σ^2_ε, respectively. Moreover, we assume that the γ_{ij} satisfy the side condition $\sum_i \gamma_{ij} = 0$, for given j.

Various sums of squares in the model (17.7.8) are again defined in the same manner as in the fixed-effects model, but there is a difference in the E(MS) column. For example, we have

$$\begin{aligned} E(SS_A) &= brE(\sum_i (\overline{y}_{i\cdot\cdot} - \overline{y}_{\ldots})^2) \\ &= brE(\sum_i (\mu + \alpha_i + \overline{\beta}_\cdot + \overline{\gamma}_{i\cdot} + \overline{\varepsilon}_{i\cdot\cdot} - \mu - 0 - \overline{\beta}_\cdot - 0 - \varepsilon_{\ldots})^2) \end{aligned}$$

Since from above we have $\sum_i \alpha_i = 0$ and $\sum_i \gamma_{ij} = 0$ for given j.

$$\overline{\beta}_{\cdot} = \frac{1}{b}\sum_j \beta_j, \quad \overline{\gamma}_{i\cdot} = \frac{1}{b}\sum_j \gamma_{ij}, \quad \overline{\varepsilon}_{i\cdot\cdot} = \frac{1}{br}\sum_j\sum_k \varepsilon_{ijk}, \quad \text{and} \quad \overline{\varepsilon}_{\cdots} = \frac{1}{abr}\sum_i\sum_j\sum_k \varepsilon_{ijk}$$

Because the expected value of cross-product terms is zero, we have

$$E(SS_A) = brE\left(\sum_i \alpha_i^2 + \sum_i \overline{\gamma}_{i\cdot}^2 + \sum_i (\overline{\varepsilon}_{i\cdot\cdot} - \overline{\varepsilon}_{\cdots})^2\right) \tag{17.7.9}$$

We use the side conditions to obtain

$$\sum_i \overline{\gamma}_{i\cdot}^2 \sim \frac{\sigma_\gamma^2}{b}\chi_{a-1}^2, \quad \sum_i (\overline{\varepsilon}_{i\cdot\cdot} - \overline{\varepsilon}_{i\cdots})^2 \sim \frac{\sigma_\varepsilon^2}{br}\chi_{a-1}^2 \tag{17.7.10}$$

Thus, we have

$$E(MS_A) = E(SS_A/(a-1)) = \sigma_\varepsilon^2 + r\sigma_\gamma^2 + \frac{br}{a-1}\sum \alpha_i^2 \tag{17.7.11}$$

Similarly, we can show that

$$E(MS_B) = \sigma_\varepsilon^2 + ar\sigma_\beta^2, \quad E(MS_{AB}) = \sigma_\varepsilon^2 + r\sigma_\gamma^2, \quad E(MS_E) = \sigma_\varepsilon^2$$

These results are summarized in Table 17.7.2.

The hypotheses of general interest are

$$H_0 : \sigma_\gamma = 0 \text{ versus } H_1 : \sigma_\gamma > 0 \tag{17.7.12}$$

$$H_0 : \text{all } \alpha_i = 0 \text{ versus } H_1 : \text{all } \alpha_i \neq 0 \tag{17.7.12a}$$

$$H_0 : \sigma_\beta = 0 \text{ versus } H_1 : \sigma_\beta > 0 \tag{17.7.12b}$$

Table 17.7.2 ANOVA table for a two-way experimental design with mixed effects

Source	DF	SS	MS	$E(MS)$
A	$(a-1)$	$br\sum(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2$	$MS_A = \frac{SS_A}{a-1}$	$\sigma_\varepsilon^2 + r\sigma_\gamma^2 + \frac{br}{a-1}\sum\alpha_i^2$
B	$(b-1)$	$ar\sum(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2$	$MS_B = \frac{SS_B}{b-1}$	$\sigma_\varepsilon^2 + ar\sigma_\beta^2$
AB	$(a-1)(b-1)$	By subtraction	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$\sigma_\varepsilon^2 + r\sigma_\gamma^2$
Error	$ab(r-1)$	$\sum\sum\sum(y_{ijk} - \overline{y}_{ij\cdot})^2$	$MS_E = \frac{SS_E}{ab(r-1)}$	σ_ε^2
Total	$abr-1$	$\sum y_{ijk}^2 - abr\overline{y}_{\cdots}^2$		

Under the hypothesis H_0 in (17.7.12), the variance ratio MS_{AB}/MS_E is distributed as $F_{(a-1)(b-1),\ ab(r-1)}$. Thus, the hypothesis $\sigma_\gamma = 0$ is rejected at the α level of significance if the ratio $MS_{AB}/MS_E > F_{(a-1)(b-1),ab(r-1);\alpha}$. Testing of the hypotheses in (17.7.12a) and (17.7.12b) depends on whether or not H_0 in (17.7.12) is rejected. Thus, if $H_0 : \sigma_\gamma = 0$ is not rejected, the hypotheses in (17.7.12a) and (17.7.12b) are tested using either the statistics MS_A/MS_{AB} and MS_B/MS_E or the statistics MS_A/MS_E and MS_B/MS_E. However, if the hypothesis $H_0 : \sigma_\gamma = 0$ in (17.7.12) is rejected, then the hypotheses H_0 in (17.7.12a) and (17.7.12b) are tested using statistics MS_A/MS_{AB} and MS_B/MS_E, respectively. The mixed-effects model can easily be extended when involved with a n-way experimental design.

17.7.3 Nested (Hierarchical) Designs

In the previous sections, we have considered experiments where every level of each factor occurs with every level of other factor; that is, the factors were *completely crossed*. We now consider experiments with two factors, say A and B, when a few levels of B occur with the first level of A, a few other levels of B with the second level of A, and so on. Such an experimental design is called a *nested* or *hierarchical design*, and we say that the factor B is *nested* in A. As an example, consider an experiment where a different doses of a drug are administered to a set of s patients in such a way that one dose level is administered to a set of s_1 patients, another dose level to a set of s_2 different patients, and so on, with $\sum_{i=1}^{a} s_i = s$. In this experiment, we say that the patients are *nested* in the drug. In another example, we might consider the case of an industrial experiment where a machines are being tested by s operators in such a way that s_1 operators work with one machine, another s_2 operators work with the second machine, and so on, with $\sum_{i=1}^{a} s_i = s$. In this experiment the operators are *nested* in the machines.

Let y_{ijk} denote the *k*th observation that is generated when the *j*th level of factor B is used within the *i*th level of A. Then, the model used for such an experiment is the following:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{k(ij)}, \quad i = 1, 2, \ldots, a;\ j(i) = 1, 2, \ldots, b;\ k = 1, 2, \ldots, r \tag{17.7.13}$$

where α_i is the effect of the *i*th level of A, $\beta_{j(i)}$ is the effect of the *j*th level of B when it is used within the *i*th level of A, and $\varepsilon_{k(ij)}$ is the random error. Note that the model (17.7.13) does not contain any interaction term. This is because the factors are not completely crossed, so there are not $a \times s \times r$ observations in a nested design experiment (see Example 17.7.1). The subscript $j(i)$ simply means that the *j*th level of B occurs within the *i*th level of A $(j = 1, 2, \ldots, s_i)$ and that $k(ij)$ denotes the error on the *k*th observation generated using the *i*th level of A and *j*th level of B, so that the *k*th replication is nested within the treatment combination of the *i*th level of A and the *j*th level of B. Generally, we assume that α_i is a fixed effect and $\beta_{j(i)}$ is a random effect. Further, it is assumed that $\sum \alpha_i = 0$ and that $\beta_{j(i)}$ and $\varepsilon_{k(ij)}$ are independently and normally distributed with mean zero and variance $\sigma_\beta^2\ and\ \sigma_\varepsilon^2$, respectively. Also, for the sake of simplicity, we assume that $s_1 = s_2 = \cdots = s_a = b$.

The various sums of squares for a nested design are defined as follows:

$$SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} y_{ijk}^2 - \frac{T_{\cdots}^2}{abr} \tag{17.7.14}$$

$$SS_A = br\sum_{i=1}^{a}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2 = \sum_{i=1}^{a}\frac{T_{i\cdot\cdot}^2}{br} - \frac{T_{\cdots}^2}{abr} \tag{17.7.14a}$$

$$SS_{B(A)} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\cdot}^2}{r} - \frac{1}{br}\sum_{i=1}^{a}T_{i\cdot\cdot}^2 \tag{17.7.14b}$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \overline{y}_{ij\cdot})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}y_{ijk}^2 - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\cdot}^2}{r} \tag{17.7.14c}$$

The expected values of various sums of squares are

$$\begin{aligned}
E(SS_A) &= brE\left(\sum(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2\right) \\
&= brE\left(\sum(\mu + \alpha_i + \overline{\beta}_{\cdot(i)} + \varepsilon\cdot_{(i\cdot)} - \mu - 0 - \overline{\beta}_{\cdot(\cdot)} - \overline{\varepsilon}_{\cdot(\cdot\cdot)})^2\right) \quad \text{since}\sum\alpha_i = 0 \\
&= brE\left[\left(\sum\alpha_i^2\right) + \sum(\overline{\beta}_{\cdot(i)} - \overline{\beta}_{\cdot(\cdot)})^2 + \sum(\overline{\varepsilon}_{\cdot(i\cdot)} - \overline{\varepsilon}_{\cdot(\cdot\cdot)})^2\right]
\end{aligned}$$

where

$$\overline{\beta}_{\cdot(i)} = \frac{1}{r}\sum_j \beta_{j(i)}, \quad \overline{\beta}_{\cdot(\cdot)} = \frac{1}{ar}\sum_i\sum_j \beta_{j(i)}, \quad \overline{\varepsilon}_{\cdot(i\cdot)} = \frac{1}{br}\sum_k\sum_j \varepsilon_{k(ij)},$$

$$\overline{\varepsilon}_{\cdot(\cdot\cdot)} = \frac{1}{abr}\sum_i\sum_j\sum_k \varepsilon_{k(ij)}$$

since the expected value of all the cross-product terms vanish.

Under the normality conditions, we have

$$\sum_i(\overline{\beta}_{\cdot(i)} - \overline{\beta}_{\cdot(\cdot)})^2 \sim \frac{\sigma_\beta^2}{b}\chi_{a-1}^2, \quad \sum(\overline{\varepsilon}_{\cdot(i\cdot)} - \overline{\varepsilon}_{\cdot(i\cdot)} - \overline{\varepsilon}_{\cdot(\cdot\cdot)})^2 \sim \frac{\sigma_\varepsilon^2}{br}\chi_{a-1}^2$$

Thus

$$E(MS_A) = \sigma_\varepsilon^2 + r\sigma_\beta^2 + \frac{br}{a-1}\sum\alpha_i^2 \tag{17.7.15}$$

Similarly we can show that

$$E(MS_{B(A)}) = \sigma_\varepsilon^2 + r\sigma_\beta^2 \tag{17.7.16}$$

$$E(MS_E) = \sigma_\varepsilon^2 \tag{17.7.17}$$

We summarize these results in Table 17.7.3.

Under the normality assumption, the testing of the hypothesis that all $\alpha_i = 0$ is carried out by using the F-statistic $MS_A/MS_{B(A)}$. To test the hypothesis $\sigma_\beta = 0$, we use the test statistic $MS_{B(A)}/MS_E$.

As we can see from the discussion above, it is important to obtain the E(MS) column. Usually, the evaluation of E(MS) term is somewhat complicated; for a general method, see Hicks (1982). If in our discussion the levels of factor A were also randomly selected from a large set of possible levels (i.e., α_i being random effects) then the ANOVA table for such an experimental design would take the form of Table 17.7.4.

Table 17.7.3 ANOVA table for a nested two-factor experimental design with mixed effects

Source	DF	SS	MS	$E(MS)$
A	$(a-1)$	$SS_A = br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\dots})^2$	$MS_A = \dfrac{SS_A}{a-1}$	$\sigma_\varepsilon^2 + r\sigma_\beta^2 + \dfrac{br}{a-1}\sum_i \alpha_i^2$
$B(A)$	$a(b-1)$	$SS_{B(A)} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$	$MS_{B(A)} = \dfrac{SS_{B(A)}}{a(b-1)}$	$\sigma_\varepsilon^2 + r\sigma_\beta^2$
Error	$ab(r-1)$	$SS_E = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{ij\cdot})^2$	$MS_E = \dfrac{SS_E}{ab(r-1)}$	σ_ε^2
Total	$abr-1$	$\sum\sum\sum(y_{ijk} - \bar{y}_{\dots})^2$		

Table 17.7.4 ANOVA table for a nested two-factor experimental design with random effects

Source	DF	SS	MS	$E(MS)$
A	$(a-1)$	$SS_A = br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\dots})^2$	$MS_A = \dfrac{SS_A}{a-1}$	$\sigma_\varepsilon^2 + r\sigma_\beta^2 + br\sigma_\alpha^2$
$B(A)$	$a(b-1)$	$SS_{B(A)} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$	$MS_{B(A)} = \dfrac{SS_{B(A)}}{a(b-1)}$	$\sigma_\varepsilon^2 + r\sigma_\beta^2$
Error	$ab(r-1)$	$SS_E = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{ij\cdot})^2$	$MS_E = \dfrac{SS_E}{ab(r-1)}$	σ_ε^2
Total	$abr-1$	$\sum\sum\sum(y_{ijk} - \bar{y}_{\dots})^2$		

Referring to Table 17.7.4, under the normality assumption, testing the hypotheses that all $\sigma_\alpha = 0$ and $\sigma_\beta = 0$ is done using the F-statistics $MS_A/MS_{B(A)}$, and $MS_{B(A)}/MS_E$, respectively. We emphasize that appropriate statistics for testing hypotheses about the effects of factors A and B depend on whether the factors A and B are fixed or random as determined by the E(MS) column. We summarize in Table 17.7.5 the results of E(MS) columns when factors A and B are fixed or random.

Example 17.7.1 (Data on an operation when operators are nested in machines) *A quality control engineer at a major manufacturing company wants to study the effects of machines and operators on a certain part. The company uses a large number of machines and operators to produce that part. Since it is not possible to study effects of each operator on each machine, the engineer decided to use a nested experimental design. The engineer randomly selects four machines and on each machine assigns three randomly selected operators (i.e., different groups of three operators on each machine). The three operator were assigned to each machine in a random order. The experimenter takes three samples of parts produced by each operator. The observed data in Table 17.7.6 give the percentage of defective parts produced.*

From Table 17.7.6 we have $T_{\dots} = 157.2$. In this example, operators B are nested in machines A so that the number of observations is $3 \times 4 \times 3 = 36$, and not $144 = 12 \times 4 \times 3$ (12 operators, 4 machines, 3 replications).

Table 17.7.5 $E(MS)$ columns for a nested two-factor experimental design

$E(MS)$	A Fixed, B Fixed	A Fixed, B Random	A Random, B Random
$E(MS_A)$	$\sigma_\varepsilon^2 + \frac{br}{a-1}\sum_i \alpha_i^2$	$\sigma_\varepsilon^2 + r\sigma_\beta^2 + \frac{br}{a-1}\sum_i \alpha_i^2$	$\sigma_\varepsilon^2 + r\sigma_\beta^2 + br\sigma_\alpha^2$
$E(MS_{B(A)})$	$\sigma_\varepsilon^2 + \frac{r}{a(b-1)}\sum_j\sum_i \beta_{j(i)}^2$	$\sigma_\varepsilon^2 + r\sigma_\beta^2$	$\sigma_\varepsilon^2 + r\sigma_\beta^2$
$E(MS_E)$	σ_ε^2	σ_ε^2	σ_ε^2

Table 17.7.6 Percentage of defective parts produced

	M_1			M_2			M_3			M_4		
	$O_{1(1)}$	$O_{2(1)}$	$O_{3(1)}$	$O_{4(2)}$	$O_{5(2)}$	$O_{6(2)}$	$O_{7(3)}$	$O_{8(3)}$	$O_{9(3)}$	$O_{10(2)}$	$O_{11(4)}$	$O_{12(4)}$
	4.6	4.7	4.2	4.1	4.0	3.9	4.8	4.5	4.1	4.2	4.1	4.9
	4.2	4.5	3.9	4.3	4.5	4.3	4.5	4.0	3.9	3.9	5.0	4.8
	4.0	4.1	4.2	4.6	4.2	4.1	4.2	4.9	4.2	4.6	5.2	5.0
$T_{ij\cdot}$	12.8	13.3	12.3	13.0	12.7	12.3	13.5	13.4	12.2	12.7	14.3	14.7
$T_{i\cdot\cdot}$	38.4			38.0			39.1			41.7		

Now the various sums of squares for the data in Table 17.7.6 are

$$SS_A = \sum \frac{T_{i\cdot\cdot}^2}{br} - \frac{T_{\cdots}^2}{abr} = 687.36 - 686.44 = 0.92$$

$$SS_{B(A)} = \sum \frac{T_{ij\cdot}^2}{r} - \sum \frac{T_{i\cdot\cdot}^2}{br} = 688.70 - 687.36 = 1.34$$

$$SS_E = \sum y_{ijk}^2 - \sum \frac{T_{ij\cdot}^2}{r} = 691.06 - 688.70 = 2.36$$

$$SS_{tot} = \sum y_{ijk}^2 - \frac{T_{\cdots}^2}{abr} = 691.06 - 686.44 = 4.62$$

These results are summarized in Table 17.7.7.

The required $\alpha = 0.05$ points are $F_{3,8;0.05} = 4.07$ and $F_{8,24;0.05} = 2.36$,, which are greater than the corresponding observed values of the F-ratios. From this analysis, we can say that based on these data, we do not reject the null hypothesis of no variation among

Table 17.7.7 ANOVA table for the data in Table 17.7.6

Source	DF	SS	MS	$E(MS)$	F-ratio
A	3	0.92	0.307	$\sigma_\varepsilon^2 + 3\sigma_\beta^2 + 9\sigma_\alpha^2$	1.83
$B(A)$	8	1.34	0.168	$\sigma_\varepsilon^2 + 3\sigma_\beta^2$	1.70
Error	24	2.36	0.099	σ_ε^2	
Total	35	4.62			

machines and no operator variation within machines. The estimates of different variance components are

$$\begin{aligned}\hat{\sigma}^2_\varepsilon &= MS_E = 0.099,\ \hat{\sigma}^2_\beta = \frac{MS_{B(A)}-MS_E}{r} = \frac{0.168-0.099}{3} = 0.023\\ \hat{\sigma}^2_\alpha &= \frac{MS_A-MS_{B(A)}}{br} = \frac{0.307-0.168}{9} = 0.015\end{aligned}$$

We have that the sum of estimates $\sigma^2_\varepsilon, \sigma^2_\beta$, and σ^2_α is

$$\text{Total} = 0.099 + 0.023 + 0.015 = 0.137$$

Thus, the contributions of various variance components are as follows:

$$\text{Machines} = 0.015/0.137 = 11.31\%, \quad \text{Operators} = 0.023/0.137 = 17.05\%,$$

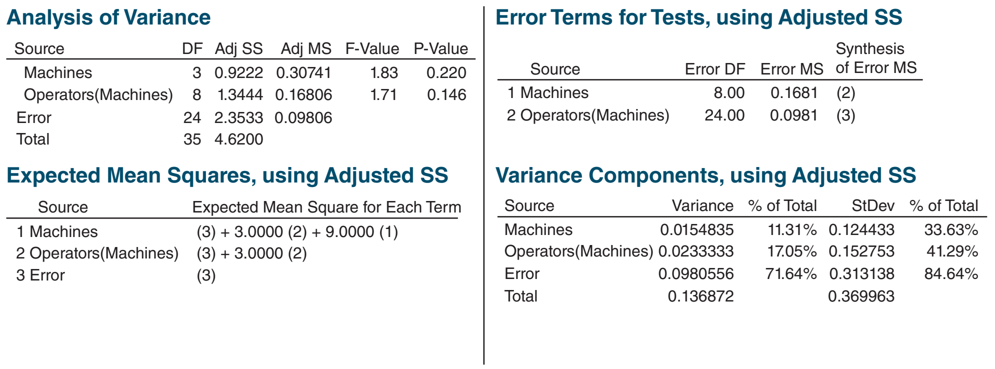
$$\text{Error} = 0.099/0.137 = 71.64\%$$

The same experimental designs can easily be extended to more than two factors when each factor is nested within the preceding one. Experiments where each factor is nested within the preceding one are called *completely nested* or *fully nested experiments*. These experimental designs are beyond the scope of this book. The interested reader is referred to Dean and Voss (1999), Montgomery (2009a,b), and other references cited at the end of this book.

Example 17.7.2 *Analyze the data in Table 17.7.6 using MINITAB and R.*

MINITAB

1. Enter the data in exactly the same manner as in the two-way experimental design. However, the identifier column corresponding to the factor that is nested must be preceded by the factor that is nesting that factor. For instance, in the present example Operators are nested in Machines so that we must enter *Machines* in C2 and *Operators* in C3.
2. Then from the Menu bar select **Stat** > **Anova** > **General Linear Model** > **Fit General Linear Model:**.
3. A dialog box appears in which you type Obs in the box below **Responses** and Machines and Operators in the box below **Factors**. Click **Random/Nest . . .** then under the **Nesting:** option type Machines in the box next Operators as Operators nested within the Machines and change the **Factor type:** to Random for both the factors. Then click **OK**. A part of the MINITAB output appears in the Session window as shown below:



```
Analysis of Variance

Source               DF  Adj SS   Adj MS  F-Value  P-Value
Machines              3  0.9222  0.30741     1.83    0.220
Operators(Machines)   8  1.3444  0.16806     1.71    0.146
Error                24  2.3533  0.09806
Total                35  4.6200

Expected Mean Squares, using Adjusted SS

Source                 Expected Mean Square for Each Term
1 Machines             (3) + 3.0000 (2) + 9.0000 (1)
2 Operators(Machines)  (3) + 3.0000 (2)
3 Error                (3)
```

```
Error Terms for Tests, using Adjusted SS

                                                Synthesis
Source                   Error DF   Error MS   of Error MS
1 Machines                   8.00     0.1681   (2)
2 Operators(Machines)       24.00     0.0981   (3)

Variance Components, using Adjusted SS

Source                  Variance  % of Total     StDev  % of Total
Machines               0.0154835      11.31%  0.124433      33.63%
Operators(Machines)    0.0233333      17.05%  0.152753      41.29%
Error                  0.0980556      71.64%  0.313138      84.64%
Total                   0.136872              0.369963
```

Clearly, all these results match those in Table 17.7.7. Note that the "Variance Component" entries are estimates of σ_α^2, σ_β^2, and σ_ε^2, respectively. In the "Expected Mean Squares" portion, σ_α^2, σ_β^2, and σ_ε^2 are denoted by (1), (2), and (3), respectively.

USING R

Solution: The R functions 'lmer()' in library 'lme4' and 'aov()' can be used to run a nested design as shown in the following R-code. The function 'lmer()' provides the variance components and the function 'aov()' provides the basic information to construct the ANOVA table and hypothesis tests. You may have to install the R package 'lme4' (install.packages('lme4')).

```
Obs=c(4.6,4.2,4,4.7,4.5,4.1,4.2,3.9,4.2,4.1,4.3,4.6,4,4.5,4.2,3.9,4.3,4.1,4.8,
4.5, 4.2,4.5,4,4.9,4.1,3.9,4.2,4.2,3.9,4.6,4.1,5,5.2,4.9,4.8,5)
Machi = c(1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4)
Oper = c(1,1,1,2,2,2,3,3,3,1,1,1,2,2,2,3,3,3,1,1,1,2,2,2,3,3,3,1,1,1,2,2,2,3,3,3)
Operators= factor(Oper)
Machines= factor(Machi)


#Fitting the nested design with random factors 'Machines and Operators' where 'Opera-
tors' are nested within 'Machines'
library(lme4)
model1 = aov(Obs ∼ Error(Machines+Machines/Operators))
# Error() indicates random factors summary(model1)


#To obtain the variance components, note that Operators are nested
within Machines
model = lmer(Obs ∼ (1|Machines)+ (1|Machines:Operators))
# (1|·) indicates random factors
summary(model)
```

PRACTICE PROBLEMS FOR SECTION 17.7

1. Consider a two-factor (A and B) experiment where factor A is run at four levels and factor B at five levels. The treatments are allocated in a completely random manner and each treatment is replicated three times. Suppose that the levels of factor B are random and of factor A are fixed. Determine the *EMS* column and give appropriate test statistics for testing the usual hypotheses.

2. Suppose that the proposed experiment in Problem 1 is conducted and the data collected are as given below. Analyze these data and state your conclusions. Use $\alpha = 0.05$.

	A_1			A_2			A_3			A_4		
B_1	19	20	12	12	15	17	19	20	16	13	20	17
B_2	20	16	15	13	20	17	20	20	13	16	20	20
B_3	12	14	14	20	16	14	17	12	20	20	14	13
B_4	13	20	17	16	15	12	12	16	20	13	20	17
B_5	16	13	12	16	14	19	17	12	20	18	16	14

3. To increase the whiteness of paper, certain fluorescent whitening agents (FWAs) selected randomly are used. A quality engineer decided to use three types (1, 2, 3) of FWA, applying each one to four different rolls of paper. This experiment was run using a nested design and the observed data collected was the whiteness index. Note here that both factors FWA and rolls are random. The data are given below. Analyze these data. Use $\alpha = 0.05$.

FWA	1				2				3			
Rolls	1	2	3	4	1	2	3	4	1	2	3	4
	82	68	81	85	66	84	82	80	69	66	73	67
	71	69	76	74	78	72	79	81	75	76	70	69
	77	70	85	75	84	70	85	84	79	82	72	79

4. An engineer examines the final finish of ball bearings manufactured on four machines. She plans an experiment using a nested design where each machine is run by three different operators. Two ball bearings from each operator are collected and tested for final finish. She selected four machines randomly and then 12 operators randomly, so that different operators were used on each machine. The data collected are given below. Analyze these data and state your conclusions. Use $\alpha = 0.05$.

Machines	1			2			3			4		
Operators	1	2	3	1	2	3	1	2	3	1	2	3
	83	78	85	81	86	88	89	82	83	83	89	93
	72	75	79	76	72	81	78	87	95	84	91	97

5. A medical team tests the effect of the generic drug metformin (used for lowering plasma glucose) marketed by three manufacturers. The team planned an experiment using a nested design in which the drug produced by each manufacturer is distributed to four different clinics; that is, a total of 12 clinics were randomly selected for this experiment. Further, two diabetic patients were randomly selected

from a group of patients treated by each clinic and their glycohemoglobin AIC measured (normal range is 4.5–5.7). The results are given below. Assume that patients have similar conditions for all known factors. Analyze these data and state your conclusions. Use $\alpha = 0.05$.

Manufacturer	1				2				3			
Clinic	1	2	3	4	1	2	3	4	1	2	3	4
	6.7	6.5	6.1	6.9	6.4	6.6	5.9	6.3	7.2	6.8	6.4	5.9
	6.6	7.0	6.8	7.2	6.0	7.1	6.9	6.2	6.5	7.2	6.3	6.2

6. The chair of a large engineering department wants to evaluate the teaching of his instructors. He randomly selects four instructors and asks them to teach a general course, which is a course on thermodynamics. Then, he takes a random sample of three students from each class and records the scores of the tests that the instructors gave during the semester. The scores recorded are shown below.

I_1			I_2			I_3			I_4		
$S_{1(1)}$	$S_{2(1)}$	$S_{3(1)}$	$S_{1(2)}$	$S_{2(2)}$	$S_{3(2)}$	$S_{1(3)}$	$S_{2(3)}$	$S_{3(3)}$	$S_{1(4)}$	$S_{2(4)}$	$S_{3(4)}$
87	98	86	76	85	96	89	83	77	79	95	75
77	95	81	91	94	84	78	75	80	98	89	75
90	86	89	90	81	82	96	98	82	98	81	95
98	88	81	92	95	88	98	98	75	87	87	89

(a) Prepare the ANOVA table for these data.
(b) Test the hypothesis of no variation among instructors. Use $\alpha = 0.01$.
(c) Determine what proportion of the total variation is due to instructors and what proportion is due to students.

7. Reanalyze the data of Problem 6, assuming that the department had only four instructors who could teach the course on thermodynamics.

17.8 A CASE STUDY

Machine Screw Case Study[1] (Background and data) A certain production process has three automatic screw machines that produce various parts. The shop has enough capital to replace one of the machines. The quality control department has been asked to conduct a study and recommend which machine should be replaced. It was decided to monitor one of the most commonly produced parts (an 1/8 in. diameter pin) on each of the machines and see which machine is the least stable. The data are collected over a three-day run for the particular part of interest. A suspected time-of-day effect must be accounted for: sometimes the machines do not perform as well in the morning, when they are started up,

[1] Source: NIST/SEMATECH Engineering Statistics Handbook, http://www.itl.nist.gov/div898/handbook, June 2003.

as later in the day. To account for this, data were collected both in the morning and in the afternoon. The data collected in Table (17.8.1) are available on the book website: www.wiley.com/college/gupta/statistics2e (Legend: Machine (1–3), Day (1–3), Time (AM: 1, PM: 2), Sample (1–10), Diameter (inch)).

Treat these data as a result of a three-factor experiment, that is, Machine, Day, and Time as factors, having 3, 3, and 2 levels, respectively.

(a) Construct the ANOVA table for the data in Table 17.8.1.
(b) Write down all the hypotheses of interest and test them using $\alpha = 0.05$.
(c) If any of the hypotheses in (b) is rejected, estimate the corresponding effects.

17.9 USING JMP

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

Assume in Problems 1–7 that one-way experimental designs are appropriate and their underlying assumptions are met

1. Four feeds are fed to pigs for a certain period of time, and at the end of this period, the pigs' weight gains in pounds are recorded as given below. Construct the ANOVA table and test the hypothesis of equal feed means at the 1% level of significance. If the null hypothesis is rejected, then use the Tukey method to detect which feed effects are significantly different.

Feeds	1	2	3	4
	57	61	41	112
	51	92	75	107
Observations	63	97	79	152
	34	81	81	85
	49	67	86	137

2. Five machines were used to manufacture ball bearings of the same size. The following data give the percentage of defective ball bearings produced by each machine on four consecutive days, where it is assumed that days have no significant effects.

	Machines	1	2	3	4	5
	1	17	11	15	13	17
Days	2	15	17	12	15	13
	3	13	19	14	16	15
	4	18	21	17	18	14

(a) Construct an ANOVA table and test the null hypothesis H_0 at the 1% level of significance that there is no differences among machines.

(b) If the null hypothesis H_0 in (a) is rejected, estimate the machine effects and use the Bonferroni method to determine which machines are different. Use $\alpha = 0.05$.

(c) Reanalyze these data, if the days have possibly significant effects.

3. In Problem 2, if the machine effects are significant, then use the S-method and the T-method to find 95% confidence intervals for the contrasts $\beta_1 - 2\beta_2 + \beta_3 - \beta_4 + \beta_5$, and $\beta_1 - \frac{1}{4}(\beta_2 + \beta_3 + \beta_4 + \beta_5)$, where $\beta_1, \ldots, \beta_5$ are the effects of the machines.

4. Sixteen determinations of the ratio of iodine to silver in four different silver preparations were made, with the results.

Preparation A	Preparation B	Preparation C	Preparation D
1.17642	1.17644	1.17643	1.17645
1.17643	1.17644	1.17642	1.17645
1.17644	1.17645	1.17644	1.17645
1.17644	1.17645	1.17646	1.17646

(a) Construct the ANOVA table for these data.

(b) Test the null hypothesis H_0 that effects due to the different preparations are not significantly different. If H_0 is rejected, estimate the four preparation effects. Use $\alpha = 0.05$.

5. Determinations were made of the yield of a chemical using three catalytic methods I, II, and III, with the results shown below (from Fraser, 1958):

Method I	Method II	Method III
47.2	50.1	49.1
49.8	49.3	53.2
48.5	51.5	51.2
48.7	50.9	52.8
		52.3

(a) Test the null hypothesis H_0 that effects due to the different catalytic methods are all zero.

(b) If in part (a) H_0 is rejected, estimate the effects. Find the p-value for the F-statistics.

6. The following data show measurements made by Heyl of the gravitational constant G for balls of gold, platinum, and glass:

Gold	Platinum	Glass
6.683	6.661	6.678
6.681	6.661	6.671
6.676	6.667	6.675
6.678	6.667	6.672
6.679	6.664	6.674
6.672		

Test the hypothesis that gold, platinum, and glass all have the same gravitational constant. If they do not have the same gravitational constant, estimate the effects due to the three materials. Find the p-value for the F-statistic.

7. Among the classrooms in the public schools of a given city, there are 12 different lighting techniques. Each of these techniques is supposed to provide the same level of illumination. To determine whether the illumination is uniform, the data shown below were compiled using four of the techniques. The classrooms are known to be homogeneous, and hence, can be discounted as a possible source of variability. Observations are in foot-candles (= 12.57 lm) on the desk surface.

Lighting techniques			
1	2	3	4
31	31	34	37
38	34	35	34
38	27	39	27
33	27	35	32
31	29	30	26

(a) Construct the ANOVA table and test the hypothesis of zero lighting-technique effects. If there are significant effects, estimate them. Find the p-value for the F-statistic.

(b) If the hypothesis of zero lighting-technique effects in (a) is rejected, then use Tukey's test to perform pairwise multiple comparisons test on treatment effects.

Assume in Problems 8–16 that RCB experimental designs are appropriate designs and their underlying assumptions are met. If the blocking factor is not indicated in the problem, then state which factor is the blocking factor.

8. An engineer wishes to test the tread wear of four brands of passenger cars tires. She conducts this experiment using five cars by randomly assigning one tire of each brand on one of the wheels of each car. Because there may be variability from one car to another, a RCB design is an appropriate design for the experiment, with cars as blocks,

and different tire brands as the treatments. After driving 10.000 miles, the amount of tread wear is measured (in mm) and recorded as shown below.

	Cars	1	2	3	4	5
Tire brand	1	2.44	2.06	2.33	2.53	2.21
	2	2.18	2.97	2.46	2.17	2.72
	3	2.45	2.54	2.36	2.56	2.73
	4	2.78	2.81	2.86	2.47	2.46

Construct the ANOVA table and analyze these data. State your conclusions. Use $\alpha = 0.05$.

9. In Problem 8, use a t-test to compare the mean tread wear for the tires of brand 1 and 4. Use $\alpha = 0.05$.

10. In Problem 8, use the Bonferroni method to perform a pairwise multiple comparisons test on treatments (tire brand) effects, and state your conclusions. Use $\alpha = 0.05$.

11. An educator believes that the learning process among low-functioning children is greatly enhanced by the teaching method used. However, she thinks that the effect of different teaching methods may vary from one age group to another, so she would like to eliminate any effect of age variation. Keeping this in mind, she conducted an experiment using five different teaching methods for four different age groups, using all teaching methods in each age group of children. All children were tested after six months of teaching and their final scores were recorded as shown below.

	Teaching method	1	2	3	4	5
Age groups	1	88	66	87	84	65
	2	84	70	63	71	69
	3	81	68	73	72	74
	4	70	67	83	87	78

Do these data provide sufficient evidence for a significant difference in mean scores among teaching methods? Use $\alpha = 0.05$.

12. To provide information on whether coatings on iron pipes are of any use, 10 pieces of pipe are each marked off in equal segments, one to receive coating A, one to receive coating B, and the third left uncoated. For each pipe, the segment to receive A, for example, is decided upon by a suitable randomization scheme. The pipes then are chosen by selecting 1 specimen from each of 10 manufacturers. Once treated, the pipes are buried in soil for a year, removed, and the depths of the corrosion pits measured. The deepest pits so found are given below:

	Coating	A	B	Untreated
	1	51	73	81
	2	41	43	52
	3	43	47	55
	4	41	53	63
Specimen	5	47	58	65
	6	32	47	50
	7	24	53	62
	8	43	38	48
	9	53	61	58
	10	52	56	59

What type of experiment is this? Construct an ANOVA table, perform tests of significance, and comment. Use $\alpha = 0.05$.

13. An experiment on flies is designed to have seven blocks of three plots each. The treatments are sprays containing 4, 8, and 16 units of an active ingredient designed to kill adult flies as they emerge from the breeding medium. The blocks comprise seven sources of the medium. Numbers of adult flies found in cases set over the plots are shown below:

Block	T_1	T_2	T_3
1	445	414	247
2	113	127	147
3	122	206	138
4	227	78	148
5	132	172	356
6	31	45	29
7	177	103	63

Test the hypothesis of no treatment effects and the hypothesis of no block effects. Use $\alpha = 0.05$.

14. In the tread-wear experiment of Problem 8, the driver of car 4 drove for a long distance with tires of brand 3 at a very low air pressure. Consequently, in the following data, the observation marked a is missing. Estimate the missing observation and reanalyze the data, using $\alpha = 0.05$.

	Cars	1	2	3	4	5
	1	2.44	2.06	2.33	2.53	2.21
Tire brand	2	2.18	2.97	2.46	2.17	2.72
	3	2.45	2.54	2.36	a	2.73
	4	2.78	2.81	2.86	2.47	2.46

15. A home-care provider wished to study the effects of different diseases on the length of home visits by her staff. She believed that the ages of the caregiver staff may also influence the length of a home visit, so the effect of age of the caregiver staff was eliminated by using an appropriate design. The data collected in this experiment are shown below:

	Disease	Cancer	Stroke	TB	Hip fracture
Age group	Under 30	38	39	47	44
	[30–40)	44	47	45	51
	[40–50)	51	48	43	49
	50 or over	55	57	63	67

Prepare an ANOVA table for these data and test the hypothesis H_0: no disease effect. Use $\alpha = 0.01$. If the null hypothesis is rejected, then use the Tukey test to perform pairwise multiple comparisons.

16. Refer to Problem 15. Suppose that during the study period three patients died, so observations on these patients were not available.

	Disease	Cancer	Stroke	TB	Hip Fracture
Age group	Under 30	38	39	47	44
	[30–40)	a	47	45	51
	[40–50)	51	48	b	c
	50 or over	55	57	63	67

Estimate the three missing observations and then reanalyze these data. Use $\alpha = 0.01$. If the null hypothesis is rejected, then use the Bonferroni method to perform a "pairwise multiple comparisons test."

Assume in Problems 17–19 that two-way experimental designs are appropriate and that their underlying assumptions are met. In addition, because of scarce resources, only one observation per cell is taken, and interactions are assumed to be zero.

17. Total solids (in %) were determined in each of six batches of wet brewer's yeast a, b, c, d, e, f by each of three analysts A, B, C, and the results obtained are given below:

Batch / Analyst	a	b	c	d	e	f
A	20.1	14.7	13.0	17.8	16.0	14.9
B	20.0	14.9	13.0	17.7	16.2	15.1
C	20.2	14.8	13.1	17.9	16.1	15.0

(a) Prepare an ANOVA table of these data.
(b) Test the null hypothesis H_0: effects due to analysts are all zero. If H_0 is rejected, estimate the effects due to the three analysts. Find the p-value for the F-statistics.
(c) Test the null hypothesis H_0': effects due to batches are all zero. If H_0' is rejected, estimate the effects. Use $\alpha = 0.05$.
(d) Assuming that batches a, b, and c are from supplier 1, and batches d, e, and f are from supplier 2, construct a contrast that can be used to test whether suppliers differ, and perform the test.
(e) Comment on the experiment.

18. During the manufacture of sheets of building material, the permeability was determined for a sheet from each of machines A, B, and C on each of nine days, with the results shown below (Hald, 1952). Perform an analysis of the data, similar to that requested in Problem 17(a), (b), (c).

	Day	1	2	3	4	5	6	7	8	9
	A	1.404	1.447	1.914	1.887	1.772	1.665	1.918	1.845	1.540
Machine	B	1.306	1.241	1.506	1.673	1.227	1.404	1.229	1.583	1.636
	C	1.932	1.426	1.382	1.721	1.320	1.633	1.328	1.689	1.703

Use the Bonferroni method for multiple comparisons among the three machines. Use $\alpha = 0.05$.

19. An experiment is conducted to test five different metals for corrosion resistance in a chemical plant environment. A secured site in the factory is chosen and a plate made from each metal is exposed for a predetermined period. At the end of that period, plates are randomly evaluated for corrosion by four analysts who measure the level of corrosion on each plate. The data obtained are shown below:

	Metals	1	2	3	4	5
	A	12.4	14.4	16.9	18.8	17.7
Analyst	B	13.0	12.4	15.0	16.7	15.2
	C	13.3	14.6	13.8	17.1	16.6
	D	14.4	12.8	13.5	18.5	17.4

(a) Prepare the ANOVA table for these data. Test the hypotheses that all metal effects are zero. Also test the hypothesis that all analyst effects are zero. Use $\alpha = 0.05$. for both tests.
(b) If any of the hypotheses in (a) are rejected, then estimate the corresponding effects.

Assume in Problems 20–23 that two-way experimental designs are appropriate designs and their underlying assumptions are met. Note that each treatment is replicated r times, $r > 1$

20. An experimenter is interested in the effects of electric shock (SH) and so-called accompanying white noise (WN) on the human galvanic skin response (sweating). Four levels of SH (0.25, 0.50, 0.75, and 1.00 mA) and two levels of WN (40 and 80 db) were selected and the following (coded) data obtained:

SH WN	0.25	0.50	0.75	1.00
40 db	3, 7,	5, 11,	9, 12,	6, 11,
	9,	13,	14,	12,
	4, 1	8, 3	11, 5	7, 4
80 db	5, 10,	6, 12,	11, 18,	7, 15,
	10,	15,	15,	14,
	6, 3	9, 5	13, 9	9. 7

(a) Prepare the ANOVA table for this set of data.
(b) Test the hypothesis that all interactions are zero. Use $\alpha = 0.05$.
(c) If the hypothesis in part (a) is not rejected, then test the hypotheses that SH effects and WN effects are zero. Find the p-value for the F-statistics you used for testing each of these hypotheses.

21. Shown below are the survival times of groups of four animals randomly allocated to three poisons and four treatments:

	Treatments	*A*	*B*	*C*	*D*
Poisons	I	0.31	0.82	0.43	0.45
		0.45	1.10	0.45	0.71
		0.46	0.88	0.63	0.66
		0.43	0.72	0.76	0.62
	II	0.36	0.92	0.44	0.56
		0.29	0.61	0.35	1.02
		0.20	0.49	0.31	0.71
		0.23	1.24	0.40	0.38
	III	0.22	0.30	0.23	0.30
		0.21	0.37	0.25	0.36
		0.18	0.38	0.24	0.31
		0.23	0.29	0.22	0.33

(a) Prepare the ANOVA table for the data of this experiment.
(b) Test the hypothesis that all interactions are zero. Use $\alpha = 0.05$.
(c) If the hypothesis in (b) is not rejected, then separately test the hypotheses that poison effects and treatment effects are zero. Use $\alpha = 0.05$.

22. Suppose the experimenter reported the data in Problem 21 by quoting the reciprocal of the observed values, obtaining the set given below.

	Treatments	A	B	C	D
	I	3.226	1.220	2.326	2.222
		2.222	0.909	2.222	1.408
		2.174	1.136	1.587	1.515
		2.326	1.389	1.316	1.613
Poisons	II	2.778	1.087	2.273	1.786
		3.448	1.639	2.857	0.980
		5.000	2.040	3.226	1.408
		4.348	0.806	2.500	2.632
	III	4.545	3.333	4.348	3.333
		4.762	2.703	4.000	2.778
		5.556	2.632	4.167	3.226
		4.348	3.448	4.545	3.030

Using this set of data, repeat the analysis and compare conclusions here with the results of Problem 21.

23. In a chemical production process, the quantity of an unwanted by-product was measured for four different catalysts (C_1, C_2, C_3, C_4) applied at three different temperatures (T_1, T_2, T_3). Measurements are expressed in percentages and the data are as shown below:

	Catalyst	C_1	C_2	C_3	C_4
	T_1	0.87	0.84	0.71	0.58
		0.79	0.81	0.68	0.64
Temperature	T_2	0.57	0.82	0.59	0.56
		0.42	0.97	0.63	0.65
	T_3	0.79	0.67	0.77	0.45
		0.76	0.73	0.71	0.59

(a) Construct a complete analysis of variance table
(b) Are the interactions zero?
(c) If the answer in (b) is yes, test the hypothesis that all catalysts effects are zero. Use $\alpha = 0.05$.
(d) Repeat (c) for temperatures at the 10% level of significance.

Assume in Problems 24–26 that two-way experimental designs are appropriate and their underlying assumptions are met. Note that each treatment is replicated r times *(r >1) and that due to nuisance variables, it was necessary to use blocking. Also assume that block effects do not interact with main effects.*

24. To test the effect of different size range of materials and extrusion pressure on wear, an experiment was conducted using two pressures (25 and 30 lb psi) and six sizes coded A, B, C, D, E, and F. Due to the time it takes to test wear, a (2×6) factorial was run on each of three days with the results as shown below. Perform an analysis of variance on these data. If the test for "no interactions" is not rejected, it may be of interest to note that the six different materials differed only in being produced using two different methods of production: specifically, A, B, and C were produced using process I and D, E, and F using process II. Is there a difference between I and II? Assume that days have no significant effects.

Day 1			Size			
Pressures	A	B	C	D	E	F
25	11.3	12.2	12.9	12.1	16.9	14.3
30	21.1	21.1	21.7	24.4	23.6	23.5
Day 2			Size			
Pressures	A	B	C	D	E	F
25	11.9	10.4	12.4	13.9	14.9	15.0
30	21.3	21.4	22.0	24.1	25.5	22.1
Day 3			Size			
Pressures	A	B	C	D	E	F
25	10.0	9.9	11.3	13.3	12.4	13.8
30	18.8	19.5	21.6	23.8	23.3	20.5

25. A (3×3) factorial experiment, carried out in a randomized block with two blocks (replications) yielded the observations given below. (The blocks correspond to different furnaces.) Calculate an appropriate analysis of variance and make the required tests of hypotheses.

	Block 1				Block 2		
	A_1	A_2	A_3		A_1	A_2	A_3
B_1	19.86	26.37	29.72	B_1	20.88	24.38	29.64
B_2	15.35	22.82	27.12	B_2	15.86	20.98	24.27
B_3	4.01	10.34	15.64	B_3	4.48	9.38	14.03

26. An experiment was designed to study the effect of varying concentrations of broth on yields of four different strains of microorganisms. Blocking was used because there are four primary sources of experimental material. The observations are shown (data from Brownlee, 1960) below:

	Strains	Concentrations		
		1	2	3
Source I	A	40	69	70
	B	52	71	91
	C	78	100	110
	D	59	76	108
Source II	A	47	76	91
	B	64	72	99
	C	73	122	143
	D	77	106	127
Source III	A	55	79	102
	B	61	83	94
	C	71	106	106
	D	78	103	127
Source IV	A	44	77	85
	B	69	75	116
	C	87	106	131
	D	76	107	125

(a) Construct a complete analysis of variance table.

(b) Separately test the hypotheses that all concentration and strain effects are zero. Use $\alpha = 0.05$.

(c) Test the hypothesis that block effects are the same at the 10% level of significance.

Assume in Problems 27–30 that Latin square designs are appropriate designs and their underlying assumptions are met.

27. In a study of gasoline consumption by city buses, four vehicles, A, B, C, D were tested. In the first run of the day over a specified course, a particular assignment of drivers a, b, c, d was used. In the next run, the drivers were reassigned to the vehicles, and so on, for all four runs, as shown in the Latin square design given below (the variable measured was (miles -10) per gallon):

	Vehicle			
Run no.	A	B	C	D
1	9.44 (a)	9.83 (b)	9.02 (c)	9.68 (d)
2	9.61 (b)	9.22 (d)	9.39 (a)	8.76 (c)
3	9.06 (d)	9.02 (c)	9.88 (b)	8.88 (a)
4	8.71 (c)	9.02 (a)	9.23 (d)	9.73 (b)

(a) Construct the ANOVA table from this data set.

(b) Test (at the 5% significance level) the null hypothesis that effects due to drivers are zero. If the hypothesis that effects of drivers are zero is rejected, estimate the four driver effects.

(c) Perform an analysis on vehicle effects similar to that in (b) for driver effects.

28. An experiment was conducted to study preconditioning of leather to determine its rate of abrasion. A large square piece of leather was cut into 36 smaller squares that were subjected to a uniform abrasion test at six humidity levels A, B, C, D, E, F in a Latin square arrangement. The rows and columns of the Latin square correspond to the two dimensions of the original large square of leather. The loss of leather due to abrasion was measured in grams. The results are shown below (from Bennett and Franklin, 1954):

	1	2	3	4	5	6
a	(C) 7.38	(D) 5.39	(F) 5.03	(B) 5.50	(E) 5.01	(A) 6.79
b	(B) 7.15	(A) 8.16	(E) 4.96	(D) 5.78	(C) 6.24	(F) 5.06
c	(D) 6.75	(F) 5.64	(C) 6.34	(E) 5.31	(A) 7.81	(B) 8.05
d	(A) 8.05	(C) 6.45	(B) 6.31	(F) 5.46	(D) 6.05	(E) 5.51
e	(F) 5.65	(E) 5.44	(A) 7.27	(C) 6.54	(B) 7.03	(D) 5.96
f	(E) 6.00	(B) 6.55	(D) 5.93	(A) 8.02	(F) 5.80	(C) 6.61

(a) Construct the ANOVA table for these data.

(b) Test (at the 5% significance level) the null hypothesis H_0 that effects due to humidity are zero. If H_0 is rejected, estimate the effects due to the six humidity levels.

29. In a 6×6 Latin square experimental design, the sums of squares corresponding to the various sources of variation are shown below. Fill out the missing columns and test the null hypothesis that (i) row effects, (ii) column effects, and (iii) treatment effects are all zero. Use $\alpha = 0.01$.

Source	Sums of squares	Degrees of freedom	Mean square	Test
Between rows	58.1			
Between columns	78.6			
Between treatments	81.3			
Error	14.2			
Total	232.2			

30. An agronomist wished to study the effects of five fertilizers (A, B, C, D, E) on the soybean crop. From past experience it is known that the fertility of land varies in two directions, so to control the effects of land fertility a 5×5 Latin square design is used. At the end of the harvest season the soybean yield is recorded and presented as coded data (yield $-$ 50 lb) that are given below:

	1	2	3	4	5
1	(A) 12	(B) 18	(C) 11	(D) 18	(E) 21
2	(B) 15	(C) 15	(D) 17	(E) 21	(A) 24
3	(C) 13	(D) 16	(E) 19	(A) 20	(B) 18
4	(D) 12	(E) 11	(A) 15	(B) 16	(C) 20
5	(E) 16	(A) 12	(B) 18	(C) 15	(D) 23

Test the null hypotheses that row effects, column effects, and treatment effects are all zero. Use $\alpha = 0.10$.

Assume in Problems 31–36 that nested experimental designs with fixed effects, mixed effects, and random effects are appropriate and their underlying assumptions are met.

31. An engineer wishes to study the tearing strength of a coated paper produced at a paper mill (tear strength is a measure of the force, applied perpendicularly to the plane of the paper, that is required to tear one or more sheets of paper clamped between two sets of jaws through a specified distance after the tear has been started, using a standard tearing tester). She takes four samples from each of the three machines used to manufacture coated paper and determines the tearing strength by taking three readings on each sample. The data obtained are shown below (M = machines, S = samples). Analyze these data and state your conclusions. Give the p-value for the F-test you used. (Here effects of machines are assumed fixed, but effects due to samples are random, etc.). Use $\alpha = 0.01$.

M_1				M_2				M_3			
$S_{1(1)}$	$S_{2(1)}$	$S_{3(1)}$	$S_{4(1)}$	$S_{1(2)}$	$S_{2(2)}$	$S_{3(2)}$	$S_{4(2)}$	$S_{1(3)}$	$S_{2(3)}$	$S_{3(3)}$	$S_{4(3)}$
28	29	30	27	19	22	24	20	30	29	33	32
22	25	32	30	21	20	22	25	33	32	35	33
29	31	28	29	22	23	21	23	35	28	34	31

32. Reanalyze the data in Problem 31, now supposing that the machines also represent a random sample selected from a large set of machines. Use $\alpha = 0.01$.

33. A rocket propellant manufacturer wishes to study the burning rate of a propellant from three production processes. Four batches of propellant are randomly selected from the output of each process, and three observations on burning rate are taken. The data obtained is shown below (from Montgomery, 2009a,b, used with permission). Analyze these data and state your conclusions. Use $\alpha = 0.01$ (P = production process, R = batch of propellant).

Production	P_1				P_2				P_3			
Propellant	$R_{1(1)}$	$R_{2(1)}$	$R_{3(1)}$	$R_{4(1)}$	$R_{1(2)}$	$R_{2(2)}$	$R_{3(2)}$	$R_{4(2)}$	$R_{1(3)}$	$R_{2(3)}$	$R_{3(3)}$	$R_{4(3)}$
	25	19	15	15	19	23	18	35	14	35	38	25
	30	28	17	16	17	24	21	27	15	21	54	29
	26	20	14	13	14	21	17	25	20	24	50	33

34. An experimenter wished to study a hereditary problem among cats. She had four sires of different breeds for the study, but she had the opportunity to select random samples of dams. She then designed an experiment that was a nested design. This design called for the mating of each sire with three dams (four batches of three dams each, where dams are randomly selected). Then, the tail length of three kittens from each dam is measured at the age of eight weeks. The data obtained are shown below:

Sires	S_1			S_2			S_3			S_4		
Dreams	$D_{1(1)}$	$D_{2(1)}$	$D_{3(1)}$	$D_{1(2)}$	$D_{2(2)}$	$D_{3(2)}$	$D_{1(3)}$	$D_{2(3)}$	$D_{3(3)}$	$D_{1(4)}$	$D_{2(4)}$	$D_{3(4)}$
	4.9	4.7	4.4	4.2	4.0	4.8	4.8	4.5	4.1	5.2	5.1	5.9
	4.3	4.2	4.8	4.3	4.1	4.3	4.9	4.6	4.9	5.0	5.2	5.8
	4.0	4.4	4.7	4.6	4.2	4.1	4.7	4.9	4.8	4.9	5.6	5.0

(a) Prepare the ANOVA table for these data.
(b) Test the hypothesis that the different breeds of sires have no effect on tail length. Use $\alpha = 0.01$.
(c) Test the hypothesis that dams within sires have the same effect on tail length. Use $\alpha = 0.05$.

35. The following data provides the plasma epinephrine obtained in 12 mice (3 random samples of 4 mice each) during 3 types of anesthesia (A_1, A_2, A_3) randomly selected from various types of anesthesia. Since only a small quantity of each type of anesthesia was available, an experiment was designed as a nested design to carry out the enquiry. Two observations are taken on each mouse:

A_1				A_2				A_3			
$M_{1(1)}$	$M_{2(1)}$	$M_{3(1)}$	$M_{4(1)}$	$M_{1(2)}$	$M_{2(2)}$	$M_{3(2)}$	$M_{4(2)}$	$M_{1(3)}$	$M_{2(3)}$	$M_{3(3)}$	$M_{4(3)}$
0.54	0.89	0.52	0.45	0.92	0.43	0.87	0.72	0.78	0.65	0.48	0.74
0.79	0.47	0.73	0.99	0.50	0.76	0.77	0.55	0.66	0.69	0.89	0.46

(a) Prepare the ANOVA table for these data.
(b) Test the hypothesis that the different types of anesthesia have the same effect. Use $\alpha = 0.05$.
(c) Test the hypothesis that mice within an anesthesia have the same plasma epinephrine. Use $\alpha = 0.05$.

36. Suppose that for the analysis in Problem 35, the factors anesthesia and mice are fixed. Reanalyze the data and draw your conclusions. Use $\alpha = 0.05$.

37. The following data shows measurements of resting heart rates of males for four age groups:

Heart rate samples

18–24	25–29	30–34	35–40
47	60	55	49
45	58	43	43
53	62	42	43
63	63	50	49
62	67	60	55
68	64	45	43
51	62	44	51
48	55	60	50
61	61	50	43
40	49	53	53

Test the hypothesis that the four age groups all have the same mean heart rate. If the hypothesis of equality is rejected, estimate the effects due to the groups. Find the p-value for the F-statistic. Assume that a *one-way* experimental design is appropriate and the underlying assumptions are met. Use $\alpha = 0.05$.

38. Analyze the data in Problem 2 using the Kruskal–Wallis nonparametric test. Use $\alpha = 0.05$.

39. Analyze the data in Problem 4 using the Kruskal–Wallis nonparametric test. Use $\alpha = 0.05$.

40. Analyze the data in Problem 5 using the Kruskal–Wallis nonparametric test. Use $\alpha = 0.05$.

41. Analyze the data in Problem 8 using the Friedman nonparametric test. Use $\alpha = 0.05$.

42. Analyze the data in Problem 13 using the Friedman nonparametric test. Use $\alpha = 0.05$.

43. Analyze the data in Problem 15 using the Friedman nonparametric test. Use $\alpha = 0.05$.

Chapter 18

THE 2^k FACTORIAL DESIGNS

The focus of this chapter is the development of special experimental designs involving factors having two levels each.

Topics Covered

- The factorial designs
- The 2^k factorial designs
- Unreplicated 2^k factorial designs
- Blocking the 2^k factorial designs
- Confounding in the 2^k factorial designs
- Yates' algorithm for the 2^k factorial designs
- The 2^k fractional factorial designs
- One-half replication of a 2^k factorial design
- One-quarter replication of a 2^k factorial design

Learning Outcomes

After studying this chapter, the reader will be able to

- Design and conduct special kinds of experiments in engineering or other scientific fields involving two or more factors, when the factors are available at two levels.
- Create blocking appropriately to avoid any kind of confounding.
- Create blocking using appropriate interactions when experiments are designed using only one-half or one-quarter replication of a 2^k design.
- Analyze data coming out of experiments employing some special techniques.
- Perform residual analysis to check the adequacy of the models under consideration.
- Summarize and interpret the results of these experiments.
- Use statistical packages MINITAB, R, and JMP to analyze the data in these experiments.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

18.1 INTRODUCTION

We continue our discussion of factorial designs that began in Chapter 17. In this chapter, we consider an important class of factorial designs in which each of the k factors has only two levels. Since these designs have exactly 2^k treatments, they are usually referred to as 2^k *factorial designs.* These designs find widespread applications in industrial environments, such as the pharmaceutical, biomedical, and chemical industries. The 2^k factorial designs are used extensively to study another important class of designs called *response surface designs.* We discuss those designs in Chapter 19.

18.2 THE FACTORIAL DESIGNS

As we noted in Chapter 17, a factorial design is a design constructed by taking all combinations of l_1 levels of factor A_1 with the l_2 levels of factor A_2, with the l_3 levels of factor $A_3, \ldots,$ and, finally, with the l_k levels of factor A_k. The complete factorial design then contains a total of $t = l_1 \times l_2 \times l_3 \times \cdots \times l_k$ "treatments" with r_j observations per treatment, $j = 1, 2, \ldots, t$. (Usually but not necessarily, a fixed number of *replicate* observations, say r, is taken on each treatment, providing a total of $N = r \times l_1 \times l_2 \times l_3 \times \cdots \times l_k$ experiments.) If $r_j = r$ for all j, then we say that we have replicated the experiment r number of times and the data obtained in this manner are called *balanced data.* Ideally, the entire program of experiments is run in a random sequence.

As an example, suppose an experimenter wants to study the effects of temperature at three levels, two types of catalyst, and four pump speeds upon the yield of a particular hydrocarbon in a fluid bed reactor. The factorial design needed here would then consist of $3 \times 2 \times 4 = 24$ treatments. To provide a measure of experimental error, each treatment might be performed twice to give a total of $N = 2 \ \times 2 \times 3 \times 4 = 48$ experiments. The experiments are then run in random order.

A primary concern of the experimenter rests in comparisons between the various treatment means. In this example, he or she may wish to compare the mean performance of the two catalysts or to determine whether there is a meaningful linear or quadratic trend associating process yield with the various temperature levels at the various pumping speeds. To estimate these and other effects, we can use certain statistics called *contrasts*, which we have defined in Chapter 17, and now discuss them in the present context.

Consider a collection of n observations Y_u, $u = 1, 2, \ldots, n$. A *contrast* of the Y_u is a linear combination of the observations of the form $\sum_{u=1}^{n} d_u Y_u$, subject to the constraint that $\sum_{u=1}^{n} d_u = 0$. Similarly, we may define contrasts between t treatment averages $\bar{Y}_j$ or treatment totals T_j, $j = 1, 2, \ldots, t$, each based on r observations by

$$\sum_{j=1}^{t} c_j \bar{Y}_j \quad \text{or} \quad \sum_{j=1}^{t} c_j T_j, \quad \text{where} \quad \sum_{j=1}^{t} c_j = 0 \tag{18.2.1}$$

Further, two contrasts $\sum_{u=1}^{n} d_u Y_u$ and $\sum_{u=1}^{n} d_u' Y_u$ are *orthogonal* if $\sum_{u=1}^{n} d_u d_u' = 0$.

Given the constraints on the constants d_u, it is possible to construct a set of $m = (n-1)$ orthogonal contrasts. Now, given a set of $n-1$ orthogonal contrasts, $L_v = \sum_{u=1}^{n} d_{uv} Y_u$, $v = 1, 2, \ldots, n-1$ it can be shown that the corrected sum of squares in an analysis of variance table is given by

$$SS_{\text{total}} = \sum_{u=1}^{n} (Y_u - \bar{Y})^2 = \sum_{v=1}^{n-1} \left(\frac{L_v^2}{\sum d_{uv}^2} \right), \tag{18.2.2}$$

So that, interestingly, the $(n-1)$ degrees of freedom associated with the corrected sum of squares in an analysis of variance table can be partitioned into $(n-1)$ separate additive components, where each single degree of freedom component $(L_v^2/\sum_u d_{uv}^2)$ is associated with a single orthogonal contrast L_v. Further, under the assumption that the errors in observation are independent $N(0,\ \sigma^2)$ random variables, the statistic

$$T_m = \frac{\sum d_u Y_u - E(\sum d_u Y)}{\sqrt{(\sum d_u^2)\, S^2}} \sim t_m \tag{18.2.3}$$

may be used to test hypotheses concerning the expected value of a contrast. Here, as previously, t_m denotes the Student t variable with m degrees of freedom and S^2 is the usual unbiased estimator of σ^2 based on m degrees of freedom. If we wish to test the hypothesis that $E\left(\sum\ d_u Y_u\right) = 0$, at significance level α, we would thus reject the hypothesis if the observed value of T_m is such that

$$|obs\ T_m| = \left| \frac{\sum d_u Y_u}{\sqrt{(\sum d_u^2) S^2}} \right| > t_{m;\alpha/2} \tag{18.2.4}$$

and do not reject the hypothesis otherwise. Now, since $t_m^2 \sim F_{1,m}$, we see that this test is equivalent to computing the *ratio* of the contrast mean square to the residual mean square in an analysis of variance table, and referring this ratio to $F_{1,m;\ \alpha}$. The limits of the $100(1-\alpha)\%$ confidence interval estimate for $E\left(\sum\ d_u Y_u\right)$ are given by

$$\sum d_u Y_u \pm t_{m;\ \alpha/2} \sqrt{\left(\sum d_u^2\right) S^2} \tag{18.2.5}$$

where m is the number of degrees of freedom associated with S^2.

Often, in a designed experiment, each of the t treatment averages $\bar{Y}_j$ is based on the same number of observations r. Here, again, individual degree of freedom *treatment* contrasts $\sum_{j=1}^{t} c_j \bar{Y}_j$ can be constructed. Also, for treatment averages based on the same number of r observations, two treatment contrasts, $\sum_{j=1}^{t} c_j \bar{Y}_j$ and $\sum_{j=1}^{t} c_j^{'} \bar{Y}_j$ are orthogonal if $\sum_{j=1}^{t} c_j^{'} c_j = 0$ and for any set of t treatment averages we may construct $(t-1)$ orthogonal treatment average contrasts. Further, the treatment sum of squares SS_A [see Chapter 17] can be partitioned into $(t-1)$ individual degree of freedom components associated with each orthogonal contrast. Each contrast sum of squares is given by

$$\frac{(\sum c_j \bar{Y}_j)^2}{r \sum c_j^2} \tag{18.2.6}$$

Also, and importantly, if Y_j's are independent normal variables, then two orthogonal contrasts of the Y_j's are also independent normal random variables.

PRACTICE PROBLEMS FOR SECTION 18.2

1. Explain in your own words: What is a factorial design?

2. Suppose that (Y_1, Y_2, Y_3, Y_4) is random sample from a normal population. Write a contrast L in terms of (Y_1, Y_2, Y_3, Y_4).
3. Refer to Problem 2. Give a set of orthogonal contrasts. What is the maximum number of orthogonal contrasts of (Y_1, Y_2, Y_3, Y_4) that may be constructed?
4. How would you test that the expected value of an orthogonal contrast $L = \sum_i c_i Y_i$ is zero?
5. Refer to Problem 4. Outline the procedure for the testing of hypothesis at the 5% level of significance for the problem

$$H_0 : \; E(L) = 0 \quad \text{versus} \quad H_1 : \; E(L) \neq 0$$

   where $L = Y_1 + Y_2 - Y_3 - Y_4$, given that $y_1 = 11, y_2 = 9, y_3 = 17, y_4 = 13$.
6. What is a $5 \times 4^3 \times 3^2 \times 2$ factorial experiment? How many factors you are studying in this experiment? How many treatments are there in this experiment?
7. Describe in your own field of application an experiment having four factors, with each factor at two levels.

18.3 The 2^k Factorial Designs

To illustrate the preceding discussion, consider a 2^k *factorial design* consisting of the $N = 2^k$ treatments formed from all possible combinations of the two levels of all the k factors. The two levels of any factor are usually called the *lower level* and the *upper level* of that factor under the control of the experimenter. For example, if one of the factors is temperature, then the two levels might be the lower and upper temperatures 100°C and 120°C. Or if the factor is qualitative, the two levels might be catalyst A and catalyst B or, as another example, the presence of catalyst A and the absence of catalyst A. These designs are frequently called *two-level* factorials without regard to the qualitative or quantitative nature of the controlled factors.

For $k = 3$ factors, the eight treatments in a 2^3 factorial design are those displayed using three equivalent notations, as in Table 18.3.1. In the first notation, the three factors are labeled A, B, and C, and the two levels of each factor are indicated by the presence or absence of the associated lowercase letter. The symbol 1 represents that treatment in which all the factors appear at their "lower" level. In the second notation, the factors are identified as 1, 2, and 3 and their two levels by 0 and 1, respectively. Finally, in the third notation, the factors are identified as x_1, x_2, and x_3, or A_1, A_2, and A_3, and their levels by plus and minus signs. The plus and minus notation provides a convenient geometric representation of the design, the eight settings (± 1, ± 1, ± 1) defining the vertices of a cube in the three space (x_1, x_2, x_3) of the factors, the center of the cube located at the origin of the x_1, x_2, x_3 coordinate system. The plus and minus indicate the upper and lower level, respectively. The design is illustrated in Figure 18.3.1b.

In all three cases, the array of treatments, called the *treatment matrix* (a submatrix of the *design matrix*), has been written down in standard order, or "Yates's order," in honor of one of the early proponents of these designs. The experimental program is, of course, run in random order. The treatment matrices for the 2^2 and the 2^4 factorials are displayed in Table 18.3.2. These designs, when viewed geometrically, are the 2^2 vertices of a square and the 2^4 vertices of the tesseract, or cube, in four space as illustrated in Figure 18.3.1a,c, respectively. The extension for $k > 4$ should be obvious.

Table 18.3.1 2^3 Factorial design in Yates's order in three notations.

Experiment	Notation 1			Notation 2			Notation 3		
number	A	B	C	1	2	3	x_1	x_2	x_3
1		1		0	0	0	$-$	$-$	$-$
2		a		1	0	0	$+$	$-$	$-$
3		b		0	1	0	$-$	$+$	$-$
4		ab		1	1	0	$+$	$+$	$-$
5		c		0	0	1	$-$	$-$	$+$
6		ac		1	0	1	$+$	$-$	$+$
7		bc		0	1	1	$-$	$+$	$+$
8		abc		1	1	1	$+$	$+$	$+$

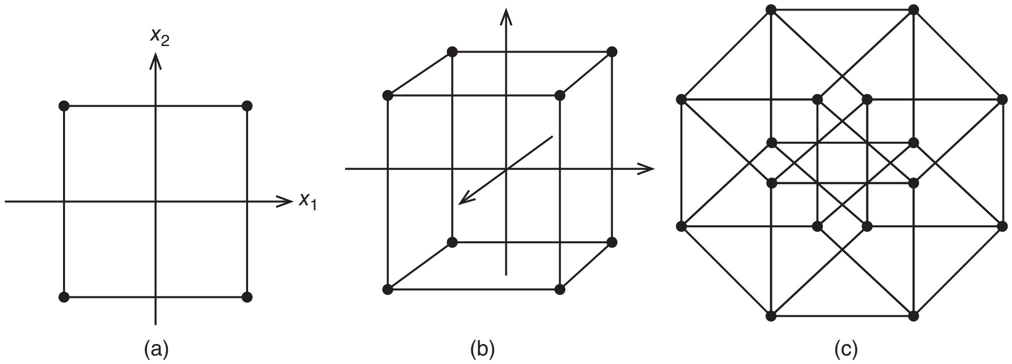


Figure 18.3.1 Geometric display of the 2^2, 2^3, and 2^4 factorial designs. (a) The 2^2 factorial, (b) the 2^3 factorial, (c) the 2^4 factorial (although one cannot "see" the tesseract, its projection to fewer dimensions can be visualized).

Each experiment is called a run, and Table 18.3.1 gives the "recipe" for the conditions of the run. For example, run 6, which generates y_6, say, is conducted by using the upper level of factor A, the lower level of factor B, and the upper level of factor C.

For the 2^k factorial experimental layout, the postulated model is

$$\begin{aligned} Y(t_1 \cdots t_k) = \mu + A_1 z_1 + A_2 z_2 + \cdots + A_k z_k + A_1 A_2 z_1 z_2 + \cdots + A_{k-1} A_k z_{k-1} z_k \\ + \cdots + A_1 A_2 \cdots A_k z_1 z_2 \cdots z_k + \varepsilon \end{aligned} \tag{18.3.1}$$

where $Y(t_1 \cdots t_k)$ is the observation taken using the treatment combination $(t_1 \cdots t_k)$, $A_1, A_2, \ldots, A_k$ are the main effects, $A_i A_j$ are the effects of the two-factor interactions, and so on; $z_i = 1$ or -1 denotes the use of the ith factor at the upper or lower level, respectively, and ε is the random error. Note from 18.3.1 that the 2^k factorial design

Table 18.3.2 Treatment matrices for the 2^2 and 2^4 factorial designs in Yates's order.

	A_1	A_2		A_1	A_2	A_3	A_4
	−	−		−	−	−	−
2^2	+	−		+	−	−	−
Factorial	−	+		−	+	−	−
	+	+		+	+	−	−
				−	−	+	−
				+	−	+	−
			2^4	−	+	+	−
			Factorial	+	+	+	−
				−	−	−	+
				+	−	−	+
				−	+	−	+
				+	+	−	+
				−	−	+	+
				+	−	+	+
				−	+	+	+
				+	+	+	+

model consists of

$$\begin{array}{ll} k & \text{Main effects} \\ k(k-1)/2 & \text{Two-factor interaction effects} \\ k(k-1)(k-2)/3\cdot 2 & \text{Three-factor interaction effects} \\ \vdots & \vdots \\ k(k-1)(k-2)\cdots(k-h+1)/h! & h\text{-Factor interaction effects} \end{array} \tag{18.3.2}$$

and a single k-factor interaction effect, for a total of $2^k - 1$ factorial effects. These are easily seen to be mutually orthogonal contrasts, and they may be estimated by setting out the plus and minus signs for each factorial effect.

Let us consider the case of the 2^3 factorial. Consulting Table 18.3.3, we see that the last column of the table simply tabulates the total of the observations taken at each of the eight treatments identified by the $(\pm, \pm, \pm)$ signs in the first three columns. Now, for example, the main effect of factor A_2 may be estimated by using the arrangement of the $(\pm)$ in the column headed $\{A_2\}$ and the column of treatment totals, as follows:

$$\hat{A}_2 = \frac{1}{r2^{3-1}}[-T_1 - T_2 + T_3 + T_4 - T_5 - T_6 + T_7 + T_8] \tag{18.3.3}$$

The factor $1/2^{3-1}$, or in general, $1/2^{k-1}$, is the reciprocal of the number of plus or minus signs in a column, while the factor r in the denominator of this expression represents the common number of observations that make up each treatment total. Hence, 18.3.3 equals

Table 18.3.3 Design matrix for the 2^3 factorial design (r observations per treatment).

Treatment (row)	A_1	A_2	A_3	A_1A_2	A_1A_3	A_2A_3	$A_1A_2A_3$	Treatment totals
1	−	−	−	+	+	+	−	T_1
2	+	−	−	−	−	+	+	T_2
3	−	+	−	−	+	−	+	T_3
4	+	+	−	+	−	−	−	T_4
5	−	−	+	+	−	−	+	T_5
6	+	−	+	−	+	−	−	T_6
7	−	+	+	−	−	+	−	T_7
8	+	+	+	+	+	+	+	T_8

$\bar{Y}_+ - \bar{Y}_-$ for factor A_2, where $\bar{Y}_+$ is average of all the $r2^{k-1}$ observations recorded when factor A_2 was at its upper level, and $\bar{Y}_-$ corresponds to average of the $r2^{k-1}$ observations taken at the lower level of factor A_2.

In general, the first k columns of signs in a contrast coefficient table, called the *design matrix* (Table 18.3.3), are identical to those given in the treatment matrix for the 2^k factorial design, and these columns may be used as bases for estimates of the k main effects. The remaining columns of signs are used to estimate "interaction" effects and are generated from the row-wise product of the signs in the first k columns.

For example, to obtain the entry in row three of the column labeled A_1A_2, we have that the entry in the column A_1 and row three is $-$, the entry in the column A_2 and row three is $+$, and hence the entry for row three in column A_1A_2 is $(-)(+) = (-)$. The generalization of this table to the 2^k factorial designs is straightforward. The estimate of the $A_iA_j \cdots A_k$ interaction effect is proportional to the sum of the products of the corresponding $\{i, j, \ldots, k\}$ elements and treatment totals $\sum_l (ij \cdots k)T_l$ and is given by

$$\text{Estimate of the } A_iA_j \cdots A_k \text{ interaction effect} = \frac{1}{r2^{k-1}} \left(\sum_{l=1}^{2^k} \{ij \cdots k\} T_l \right) \tag{18.3.4}$$

where $\{ij \cdots k\}$ stands for the 2^k elements of plus and minus signs associated with the $A_iA_j \cdots A_k$ interaction effect. Thus, in the example above, the estimate of the A_1A_2 interaction effect would be (see the A_1A_2 column of Table 18.3.3)

$$(\widehat{A_1A_2}) = \text{Estimate of } A_1A_2 = \frac{1}{r(4)}[+T_1 - T_2 - T_3 + T_4 + T_5 - T_6 - T_7 + T_8] \tag{18.3.4a}$$

The associated single degree of freedom component of the treatment sum of squares is, from (18.2.6) and (18.3.4),

$$SS\{A_iA_j \cdots A_k\} = \frac{(\sum (ij \cdots k)T_l)^2}{r2^k} = r2^{k-2}[\text{Estimate of effect}]^2 \tag{18.3.5}$$

The reader may check easily that the $(2^k - 1)$ estimates are mutually orthogonal contrasts, and that each contrast is the difference between two averages, say $\bar{Y}_+$ and $\bar{Y}_-$, of $r2^{k-1}$

observations each. (Recall the example of $\hat{A}_2$ in 18.3.3). Thus, the variance for *each* estimate of the effects is

$$Var(\bar{Y}_{+} - \bar{Y}_{-}) = Var[\text{Estimate of effect}] = \left(\frac{1}{r2^{k-1}} + \frac{1}{r2^{k-1}} \right) \sigma^2 = \frac{1}{r2^{k-2}} \sigma^2 \qquad (18.3.6)$$

We will denote an estimate of the variance in (18.3.6) by $\hat{V}$, that is,

$$\hat{V} = \frac{1}{r2^{k-2}} S^2 \qquad (18.3.6a)$$

where $S^{\,2}$ is the estimate of σ^2 obtained by dividing the appropriate error sum of squares by its associated degrees of freedom, say m. Here, S^2 is independent of the treatment effects.

A test of a hypothesis about the expected value of an effect is provided by using the fact that the following is distributed as Student's t:

$$T_m = \frac{[\text{Estimate of effect}] - E[\text{Estimate of effect}]}{\sqrt{\hat{V}}} \sim t_m \qquad (18.3.7)$$

An interval estimate for any effect is given by

$$[\text{Estimate of effect}] \pm t_{m;\ \alpha/2} \times \sqrt{\hat{V}} \qquad (18.3.8)$$

Example 18.3.1 (Preparing a colored fabric) *An experimenter wishes to determine the effects of a new dyestuff and alternative methods for preparing fabric upon fabric colorfastness. The 2^2 factorial design given in Table 18.3.4 was employed in which the upper and lower levels of factor* X_1 *correspond to the new dyestuff and standard dyestuff, respectively, and the upper and lower levels of the factor* X_2 *represent the alternative and standard methods for preparing the fabric, respectively. The measured response* Y *(here modified and coded for convenience) was the loss in reflectance of dyed fabric after 20 hours of exposure to a carbon-arc lamp, in keeping with standards in AATCC Method 16A-1864 (American Association of Textile Chemists and Colorists). To provide estimates of treatment effects with sufficient precision, it was decided to replicate the experimental program five times.*

Table 18.3.4 Data for 2^2 factorial in five replicates.

	Studied factors		Design variables		Responses y_{ij} (loss of reflectance) blocking variable: days					Treatment	
	Type of	Preparation								Totals	Averages
Treatment	dyestuff	method	A_1	A_2	(i)	(ii)	(iii)	(iv)	(v)	$T_{i.}$	$\bar{Y}_{i.}$
1	Old	Standard	$-$	$-$	23	21	23	21	22	110	22.0
2	New	Standard	$+$	$-$	30	28	31	28	23	140	28.0
3	Old	Alternative	$-$	$+$	12	8	5	9	6	40	8.0
4	New	Alternative	$+$	$+$	23	19	17	14	17	90	18.0
			Day totals $T_{.j}$		88	76	76	72	68	$T_{..} = 380$	
			Day averages $\bar{Y}_{.j}$		22.0	19.0	19.0	18.0	17.0		$\bar{Y}_{..} = 19$

It was also felt that day-to-day variability might inflate the variance of the observations. So to protect the treatment comparisons against this unwanted source of variability, the 2^2 factorial program was randomly run once on each of the five days. The 2^2 factorial design and recorded responses are given in Table 18.3.4; note that $k = 2$, $r = 5$.

Solution: First, we analyze these data using the technique of a randomized complete block design explored in Chapter 17. A test of the hypothesis that there are no treatment effects is carried out by first postulating the model (see Section 17.4)

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{18.3.9}$$

where the Y_{ij} are recorded observed reflectance measurements, $i = 1, \ldots, 4$, $j = 1, \ldots, 5$, and where the α_i are here defined to be the treatment effects, β_j the effects due to days (the blocks), and the ε_{ij} are random errors distributed independently as $N(0, \sigma^2)$. We then use the analysis of variance (see Table 17.4.2) to estimate σ^2, the variance of the observations, and finally to test the hypothesis that the treatment effects α_i are all equal to zero. We have ($r = 5$, $k = 2$, $r2^k = 20$)

$$\begin{aligned}
SS_{\text{total}} &= \sum\sum y_{ij}^2 - \frac{T_{..}^2}{r2^k} \\
&= 23^2 + 30^2 + \cdots + 17^2 - \frac{(380)^2}{20} = 8396 - 7220 = 1176 \\
SS_{treat} &= \frac{1}{r}\sum T_{i\cdot}^2 - \frac{T_{..}^2}{r2^k} \\
&= \frac{1}{5}(110^2 + 140^2 + 40^2 + 90^2) - \frac{(380)^2}{20} = 8280 - 7220 = 1060 \\
SS_{days} &= \frac{1}{2^k}\sum T_{\cdot j}^2 - \frac{T_{..}^2}{r2^k} \\
&= \frac{1}{4}(88^2 + 76^2 + \cdots + 68^2) - \frac{(380)^2}{20} = 7276 - 7220 = 56
\end{aligned}$$

These results are summarized in the ANOVA Table 18.3.5. Note that SS_E is found by subtraction, that is, $SS_E = SS_{Total} - (SS_{treat} + SS_{days})$.

Table 18.3.5 ANOVA table for the data in Table 18.3.4.

Source	*DF*	*SS*	*MS*	*E(MS)*	*F*-ratio
Treatments	3	1060	353.3	$\sigma^2 + (5/3)\sum\alpha_i^2$	70.66
Days	4	56	14.0	$\sigma^2 + \sum\beta_j^2$	2.8
Error	12	60	5.0	σ^2	
Total	19	1176			

To test the hypothesis that the treatment effects α_i are all zero, we note that under this hypothesis, both the treatment mean square and the error mean square are independent estimates of σ^2, and hence their ratio is distributed as an F-distribution with (3, 12) degrees of freedom. The observed ratio is then $F = 353.3/5.0 = 70.66$, and from Table A.7, this is found to be a rare event since the Prob$\{F_{3,12} \geq 7.2258\} = 0.005$ (here, as elsewhere in this chapter, the significance level of a test is 0.05). Hence, the hypothesis of all $\alpha_i = 0$ is rejected. We now turn to making additional inferences about the α_is.

Employing the results of Equations (18.3.4) and (18.3.5), we can separate the three degrees of freedom for treatments into individual orthogonal one-degree-of-freedom contrasts associated with (i) the main effect of factor A_1 (changing dyestuff), (ii) the main effect of factor A_2 (changing fabric preparation), and (iii) a measure of the dyestuff-preparation interaction effect A_1A_2. To facilitate the computation of these contrasts (see (18.3.4)) or, equivalently, the estimation of these effects, the contrast coefficients and treatment totals are displayed in Table 18.3.6. The estimate of the variance of an effect is $\hat{V} = [1/(5)(1)]S^2 = [1/(5)(1)](5.0) = 1.0$ (see Equation (18.3.6a) and the analysis of variance Table 18.3.5).

Table 18.3.6 Contrast coefficients for factorial effects.

Contrast coefficients			Treatment
$\{A_1\}$	$\{A_2\}$	$\{A_1A_2\}$	totals T
$-$	$-$	$+$	110
$+$	$-$	$-$	140
$-$	$+$	$-$	40
$+$	$+$	$+$	90

The estimated effects are

$$\{A_1 \text{ dyestuff effect}\} = \frac{1}{5(2^{2-1})}[-110 + 140 - 40 + 90] = 8.0$$

$$\{A_2 \text{ preparation effect}\} = \frac{1}{5(2^{2-1})}[-110 - 140 + 40 - 90] = -12.0$$

$$\{A_1A_2 \text{ interaction effect}\} = \frac{1}{5(2^{2-1})}[+110 - 140 - 40 + 90] = 2.0$$

The associated single-degree-of-freedom portions of the treatment sum of squares are given by (see Equation 18.3.5)

$$\text{Sum of squares due to } A_1 \text{ (dyestuff)} = 5(8.0)^2 = 320$$

$$\text{Sum of squares due to } A_2 \text{ (preparations)} = 5(-12.0)^2 = 720$$

$$\text{Sum of squares due to } A_1A_2 \text{ (interaction)} = 5(2.0)^2 = 20$$

These results are summarized in Table 18.3.7. It is interesting to note that these sums of squares total to 1060; that is, we have successfully partitioned the treatment sum of squares of Table 18.3.5.

A test of the hypothesis that the effect of changing dyestuff is zero, that the main effect of A_1 is zero, is provided by the ratio of the dyestuff and error mean squares (see Table 18.3.5). Thus the relevant observed F is $F = MS(A_1)/MS_E$, which under H_0: main

Table 18.3.7 Partitioning the treatment sum of squares into individual degree of freedom of components.

Source	Sum of squares	Degrees of freedom	Mean squares
A_1 (dyestuff) effect	320	1	320
A_2 (preparation) effect	720	1	720
A_1 A_2 (interaction) effect	20	1	20
Total treatment sum of squares	1060	3	

effect of A_1 is zero is such that $F \sim F_{1,12}$. But the observed F is

$$F(A_1) = \frac{320}{5} = 64.0,$$

and this is a rare event since 64.0 is much greater than 4.74, where $\text{Prob}\{F_{1,12} \geq 4.74\} = 0.05$. Thus, we reject the hypothesis that the main effect of A_1 is zero. An equivalent test is provided by using the t-statistic (see Equations (18.3.6)–(18.3.7)),

$$(obs.\ T) = \frac{8.0 - 0}{\sqrt{5.0/5(2^{2-2})}} = 8.0$$

We note, once again, that $F = (obs\ T)^2$. Similarly, the hypothesis that there is no effect due to factor A_2 (preparation) must be rejected since the observed $F(A_2) = 720/5.0 = 114.0$. However, there is no strong evidence that an interaction effect between the type of dyestuff and the method of preparation exists, since the corresponding test of the null hypothesis "interaction effect$\,=\,$0" uses the observed $F(A_1A_2) = 20/5 = 4.0$, and $\text{Prob}\{F_{1,12} > 4\} > 0.05$. The 95% confidence interval for each individual effect is (see 18.3.8 and note that $t_{12,\ .025} = 2.179$)

$$\text{Estimate of effect} \pm 2.179\sqrt{\frac{5.0}{5(2^{2-2})}} = \text{Estimate of effect} \pm 2.179$$

since the estimate of the variance of an effect is given by (see 18.3.6a)

$$\hat{V} = \frac{1}{5(2^{2-2})}\hat{\sigma}^2 = \frac{1}{5}S^2 = \frac{1}{5} \times 5 = 1$$

In summation, on the basis of the evidence provided by the 2^2 factorial, the effect of changing to the new dyestuff will be an increase in the reflectance and hence, the colorfastness of the dyed fabric, by 8.0 ± 2.2 units, whereas changing to the new method for preparing the fabric prior to dyeing has a deleterious effect of -12.0 ± 2.2 units. Changing to the new dyestuff and continuing with the standard mode of preparation of fabric is thus strongly suggested by these data.

The analysis of these data need not end with the investigation of the effects of the treatments. The experimenter could also test the hypothesis that the day-to-day effects β_j were zero. Investigating the variation in color fastness due to days is not the primary objective of the experimenter, but the design and associated analysis of variance table provide a ready opportunity to perform this test of hypothesis. The corresponding F-ratio is observed to be $F = 14.0/5.0 = 2.8$. Now under "H_0 : days have no effect," $F \sim F_{4,12}$, so $F = 2.8$ is not a rare event at the 5% level of significance since $\text{Prob}\{F_{4,12} \geq 3.26\} = 0.05$. However, it may be considered a "rare" event at the 10% level since $\text{Prob}\{F_{4,12} \geq 2.48\} = 0.10$. A variety of interpretations are now open to the experimenter. He could decide, on the basis of the test at the 5% level of significance, that no day effects existed and that blocking the experiment by days was an unnecessary nuisance. He might then *pool* the sums of squares and degrees of freedom for days and error and produce (see Table 18.3.5) the pooled estimate of variance $S^2 = (60 + 56)/(12 + 4) = 116/16 = 7.25$ with 16 degrees of freedom. We note that the hypothesis that the treatment effects are zero is still rejected for the new observed $F = 353.3/7.25 = 48.73$, an extraordinarily rare event since $\text{Prob}\{F_{3,16} \geq 6.30\} = 0.005$. Or the experimenter might decide, on the basis of a 10% significance test, that blocking to eliminate day-to-day effects had been worthwhile since real day-to-day effects were detected and use would be only made of the mean square error $S^2 = 5.0$ in the analysis. Alternatively, he or she might still pool the day-to-day error contributions to see whether the treatment effects are detectably nonzero in the presence of this additional acknowledged source of variation due to days. The fact that they are, in this example, enhancing the experimenter's ability to make statements about the treatments across future days.[1]

The reader might have noticed that there seems to be a trend across days, the daily average diminishing gradually over time. Since days are equally spaced in time, one can easily construct orthogonal single degree of freedom contrasts and associated sum of squares reflecting the day-to-day variability that could be assigned to a linear or quadratic trend across days. The necessary sets of constant coefficients for the linear and quadratic trends are displayed in Table 18.3.8. The sum of squares and degrees of freedom for days may now be partitioned as illustrated in Table 18.3.9. The successive daily averages are averages of the four observations taken on that particular day.

Table 18.3.8 Coefficients of linear and quadratic effect contrasts.

	Successive daily averages					Contrast	Contrast SS
	22.0	19.0	19.0	18.0	17.0	$\sum c_i \bar{y}_{i\cdot}$	$= r(\sum c_i \bar{y}_{i\cdot})^2 / \sum c_i^2$
Linear effects contrast	−2	−1	0	1	2	−11	$4(-11)^2/10 = 48.40$
Quadratic effects contrast	2	−1	−2	−1	2	3	$4(3)^2/14 = 2.57$

Table 18.3.9 Partitioning block sum of squares.

	SS	Degrees of freedom	Mean square	
Linear day effect	48.40	1	48.40	$F_{1,12} = 9.68$
Quadratic day effect	2.57	1	2.57	$F_{1,12} < 1$
Other effects (by subtraction)	5.03	2	2.51	$F_{2,12} < 1$
Total sum of squares for days	56.00	4	14.0	

[1] The experimenter here assumes that there is no day-by-day treatment interaction. If, before running the experiments, he had thought such interactions likely, the design would have to be modified to permit their estimation. This could be accomplished by repeating the treatments within each day (see Chapter 17).

From Table 18.3.9, we note that the other effects and quadratic effects are insignificant. Now, we conduct the test of hypothesis that no linear trend exists (i.e. the linear contrasts among the true treatment means equals zero against the alternative that the linear contrast is not zero), given by the ratio of the linear effect mean square divided by the error mean square, $S^{\,2} = 5.0$, as found in Table 18.3.5. Thus, the relevant F-statistic is observed to be $48.40/5.0 = 9.68$. Since Prob$\{F_{1,12} \geq 4.74\} = 0.05$, this observed value of 9.68 is a rare event, and we can reject the hypothesis that no linear trend exists, or, do not reject the hypothesis that a linear trend exists. We note that this apparent linear trend accounts for almost all the variation between days. In this instance, one assignable cause was the wear on the electrodes in the carbon-arc lamp used in the instrument for measuring reflectance. For linear, quadratic, and higher-order effects contrasts coefficients, the reader is referred to Hicks (1982).

The ease and richness of this analysis are due, in very large part, to the experimental design.

PRACTICE PROBLEMS FOR SECTION 18.3

1. Write treatment matrices in all three notations for the 2^4 factorials in Yates order.
2. Write a design model for a 2^4 factorial design.
3. Refer to Problem 2. How many two-factor interactions, three-actor interactions, and four-factor interactions are there in the model?
4. Refer to Problem 2. Suppose that the whole experiment is replicated three times. Give estimates of the main effects in terms of treatment totals.
5. Write the design matrix for a 2^3 factorial design.
6. Refer to Problem 5. Suppose that the whole experiment is replicated four times. Give an estimate of the three-factor interaction in terms of treatment totals.
7. Estimate the standard error of the estimates of various main effects and interactions in Problem 6.
8. Determine a 95% confidence interval for the three-factor interaction in Problem 6.
9. Determine a 99% confidence interval for the main effects in Problem 6.

18.4 Unreplicated 2^k Factorial Designs

For moderate values of k, $k \geq 4$, the total number of treatments specified by a 2^k factorial design quickly becomes large and experimenters often become unwilling to repeat, or replicate, the experimental program. When replication is ruled out, no estimate of σ^2 is available, or none exists that can be constructed from replicated observations.

However, an estimate of σ^2 can be constructed whenever the number of variables or factors k is large. To explain, when working with k factors, it is unlikely that all the $(2^k - 1)$ factorial effects will in fact be large. Under the assumption that the response being investigated changes smoothly over the range of the factors being varied, it becomes unlikely that high-order interaction effects exist.

When they exist, the magnitude of such effects is usually relatively small compared to the main effects or the lower-order interactions. In circumstances where these assumptions seem reasonable, the $[(2^k - 1) - k - k(k-1)/2]$ degrees of freedom available for the estimation of the three-factor and higher-order interaction effects are now used to provide an estimate of the variance.

In other cases it could become clear, after experiments have been performed, that h ($h < k$) of the factors have only *very small* or *no effect* on the response when compared to the effects of the remaining $(k-h)$ factors. When this occurs, the experimenter often declares the h factors (over the range studied) to have zero effects or to have effects whose magnitudes cannot be distinguished from the contribution of the random errors. The program used then becomes a 2^{k-h} factorial replicated 2^h times, providing $[(2^h-1)\times 2^{k-h}]$ degrees of freedom for estimating σ^2.

In general, the full 2^k factorial designs can be viewed as "equal opportunity" designs, since they permit the orthogonal estimation of all (2^k-1) factorial effects, each estimated with minimum variance $\sigma^2/2^{k-2}$. Often, the experimenter is anxious to determine which subset of these (2^k-1) candidate effects have the largest influence on the response. We say that the experimenter wishes to *screen* the many effects to discover the important ones. When the factorial design is not replicated, this search becomes difficult since the experimenter will have to determine which of the *estimated* effects are due to the experimental error and which are reflections of real, large effects. To help identify the real effects, the estimates of the effects may be plotted on normal probability paper. Since linear combinations of random variables are statistics tending to have a normal distribution, those estimates that have values that are primarily due to the errors of observation should look like events from a normal distribution.

Following Section 5.8, the (2^k-1) estimates of the factorial effects are first ordered, say, as $e_{(1)},\ldots,e_{(2^k-1)}$ and these ordered values $e_{(i)}$ are plotted against $P_i=(i-0.5)/(2^k-1), i=1,2,\ldots,2^k-1$, on normal probability paper. If the effects have true value zero, then the ordered estimated effects $e_{(i)}$ will, when plotted against P_i, tend to fall along a *straight line*, confirming the hypothesis that these estimates are due solely to errors. However, the estimates of largest magnitude will, *if they reflect real effects*, lie off the straight-line. Having distinguished these large estimates from estimates that may be assumed to be manifestations of error only, the experimenter may use some or all of the degrees of freedom associated with the "error-like" effects and construct an estimate of σ^2. We illustrate this method for constructing an estimate of σ^2 from an unreplicated factorial experiment in the following example.

Example 18.4.1 (Chemical yields) *An experimenter is interested in studying the effects of* k = *4 factors on the yield of a chemical where the four factors are temperature, speed of agitation, catalyst concentration, and pressure. The experiments comprising a 2*^{4} *factorial design were performed in a random sequence. The lower and upper levels of these factors are as shown in Table 18.4.1.*

Table 18.4.1 Lower and upper levels of various factors.

Factors	Lower level	Upper level
Temperature (°C) A	30.0	32.0
Speed (1000 rpm) B	1.0	1.2
Catalyst (mol) C	0.6	1.0
Pressure (100 psi) D	7.0	10.0

The yield of the chemical process $y_i, i=1,2,\ldots,16,$ *and the experimental design in two alternative notations are given in Table 18.4.2. The table of contrast coefficients required for estimating the 15 factorial effects is displayed in Table 18.4.3.*

Table 18.4.2 Yield of the chemical and design matrix in two notations.

	Equivalent design levels				Alternate designation	Observations
Run Number	A	B	C	D	$A\ B\ C\ D$	y_i
1	−	−	−	−	1	62
2	+	−	−	−	a	88
3	−	+	−	−	b	63
4	+	+	−	−	ab	83
5	−	−	+	−	c	88
6	+	−	+	−	ac	80
7	−	+	+	−	bc	99
8	+	+	+	−	abc	92
9	−	−	−	+	d	65
10	+	−	−	+	ad	123
11	−	+	−	+	bd	65
12	+	+	−	+	abd	121
13	−	−	+	+	cd	97
14	+	−	+	+	acd	105
15	−	+	+	+	bcd	92
16	+	+	+	+	$abcd$	117

Table 18.4.3 Design matrix and responses.

				Two-factor Interactions						Three-factor Interactions				Four-factor Interactions	Responses
A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	$ABCD$	y
−	−	−	−	+	+	+	+	+	+	−	−	−	−	+	62
+	−	−	−	−	−	−	+	+	+	+	+	+	−	−	88
−	+	−	−	−	+	+	−	−	+	+	+	−	+	−	63
+	+	−	−	+	−	−	−	−	+	−	−	+	+	+	83
−	−	+	−	+	−	+	−	+	−	+	−	+	+	−	88
+	−	+	−	−	+	−	−	+	−	−	+	−	+	+	80
−	+	+	−	−	−	+	+	−	−	−	+	+	−	+	99
+	+	+	−	+	+	−	+	−	−	+	−	−	−	−	92
−	−	−	+	+	+	−	+	−	−	−	+	+	+	−	65
+	−	−	+	−	−	+	+	−	−	+	−	−	+	+	123
−	+	−	+	−	+	−	−	+	−	+	−	+	−	+	65
+	+	−	+	+	−	+	−	+	−	−	+	−	−	−	121
−	−	+	+	+	−	−	−	−	+	+	+	−	−	+	97
+	−	+	+	−	+	+	−	−	+	−	−	+	−	−	105
−	+	+	+	−	−	−	+	+	+	−	−	−	+	−	92
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	117

Table 18.4.4 Estimated factorial effects using Table 18.4.3.

Main effects		Two-factor interactions		Three-factor interactions		Four-factor interactions	
A	22.25	AB	1.25	ABC	3.25	$ABCD$	1.50
B	3.00	AC	−17.75	ABD	2.50		
C	12.50	AD	14.50	ACD	2.50		
D	16.25	BC	4.50	BCD	−2.25		
		BD	−1.75				
		CD	−3.25				

Order i	Effect	Estimate	$(i-0.5)/15$
15	1	22.25	0.9677
14	4	16.25	0.9000
13	14	14.50	0.8333
12	3	12.50	0.7667
11	23	4.50	0.7000
10	123	3.25	0.6333
9	2	3.00	0.5667
8	124	2.50	0.5000
7	134	2.50	0.4333
6	1234	1.50	0.3667
5	12	1.25	0.3000
4	24	−1.75	0.2333
3	234	−2.25	0.1667
2	34	−3.25	0.1000
1	13	−17.75	0.0333

Using (18.3.4), we can easily obtain the estimated factorial effects, which are displayed in Table 18.4.4. An alternative and more rapid estimation technique for this example is provided by Yates's algorithm discussed in Section 18.5.2.

Now let us consider the problem of obtaining an estimate of σ^2. As a preliminary for determining an estimate of σ^2, the ordered effects (see Table 18.4.4) are plotted on normal paper, shown in Figure 18.4.1. On viewing this plot (and remembering that extreme points should have little weight in orientating the fitted line on normal probability paper), we note that the smallest estimates (those near zero) lie reasonably along a straight line while the largest estimates are far off the line. This suggests strongly that these large estimates are not acting in a matter compatible with the suggestion that they are due to random errors.

One immediate interpretation of the data open to the experimenter is that factor B (speed of agitation) has no influence on the response since none of the large effects involve B. Under this assumption, we drop B, and the program becomes a replicated 2^3 factorial in variables A, C, and D with $r=2$. We then may find an estimate of σ^2, from the corresponding analysis of variance table based on this simplifying assumption, as given in Table 18.4.6.

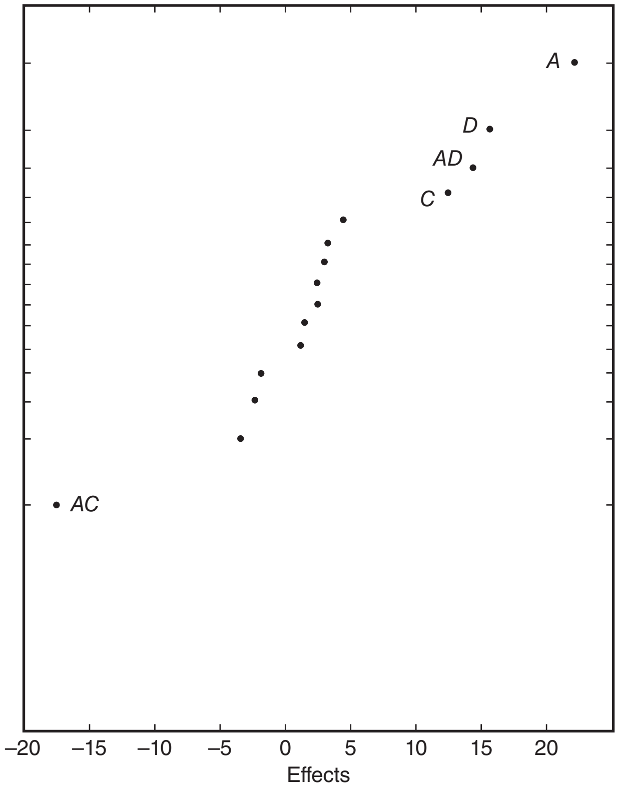


Figure 18.4.1 Ordered effects for the 2^4 factorial design of Table 18.4.4, plotted on normal probability paper.

The sum of squares contribution of the main effect of temperature (A) is calculate to be 1980.25 ($r = 2$ and $k = 3$), and see (18.3.5). This value of SS_A is arrived at as follows. Before dropping B, we consider the estimate of A using Table 18.4.3, as given by

$$\hat{A} = (-62 + 88 - 63 + 83 - \cdots - 92 + 117)/8 = 22.25,$$

and using the "Response" column of Table 18.4.5, we see that we can write $\hat{A}$ as

$$\hat{A} = (-(62 + 63) + (88 + 83) - \cdots + (105 + 117))/8 = 22.25$$

which is the recipe for the estimate of A dictated if using the design of Table 18.4.5 with $r = 2$ replicates. Hence, SS_A found from Table 18.4.5 has the same value as that obtained previously when working from Table 18.4.3, namely $r2^{k-2}(\hat{A})^2 = 4\hat{A}^2$, since now we have $r = 2, k = 3$ (note that here $k = 3$ because B is dropped).

Similarly, we can obtain all other sum of squares shown in Table 18.4.6. Using the estimate $S^2 = 29.00$ (see Table 18.4.6), based on eight degrees of freedom, the three main effects (temperature A, catalyst concentration C, and pressure D) are clearly significant, as are the temperature–catalyst and temperature–pressure interactions, since Prob$\{F_{1,8} \geq$

Table 18.4.5 The 2^3 design and observations if dropping B from the design of Table 18.4.3.

A	C	D	AC	AD	CD	ACD	Responses	Responses totals
−	−	−	+	+	+	−	62, 63	125
+	−	−	−	−	+	+	88, 83	171
−	+	−	−	+	−	+	88, 99	187
+	+	−	+	−	−	−	80, 92	172
−	−	+	+	−	−	+	65, 65	130
+	−	+	−	+	−	−	123, 121	244
−	+	+	−	−	+	−	97, 92	189
+	+	+	+	+	+	+	105, 117	222

Table 18.4.6 Analysis of variance assuming all effects containing B are zero.

Source	SS	$d.f.$	MS	F-Ratio
A	1980.25	1	1980.25	$F = 68.3$
C	625.00	1	625.00	$F = 21.6$
D	1056.25	1	1056.25	$F = 36.4$
AC	1260.25	1	1260.25	$F = 43.5$
AD	841.00	1	841.00	$F = 29.0$
CD	42.25	1	42.25	$F = 1.5$
ACD	25.00	1	25.00	$F = 0.9$
Residual	232.00	8	$29.00 = s^2$	
Total	6062	15		

$5.32\} = 0.05$. The 95% confidence limits for the effects are (see Equation 18.3.8)

$$\{\text{Estimate of effect}\} \pm t_{8;\ .025}\sqrt{\left(\left(\frac{1}{8}+\frac{1}{8}\right)S^2\right)} = \{\text{Estimate of effect}\} \pm 2.306\sqrt{\frac{29.0}{4}}$$
$$= \{\text{Estimate of effect}\} \pm 6.21 \quad (18.4.1)$$

An alternative procedure for obtaining an estimate of σ^2 is to assume that the effects of the four three-factor interactions and the single four-factor interaction are insignificant and to pool together the sums of squares associated with these high-order interactions. Consulting Table 18.4.4, we see that the sum of squares for these effects is

$$2(2^{4-2})[(3.25)^2 + (2.50)^2 + (2.50)^2 + (-2.25)^2 + (1.50)^2] = 121.5 \quad (18.4.2)$$

Thus an estimate of σ^2 based on five degrees of freedom is $S^2 = 121.5/5 = 24.3$. Inferences concerning which effects were important would not be materially changed had the experimenter arrived at the estimate of σ^2 in this fashion. In this example, as in the previous example of the replicated 2^2 factorial design (see Section 18.3), analyses can vary slightly.

PRACTICE PROBLEMS FOR SECTION 18.4

1. A study to determine whether modest changes in four critical dimensions in an automobile carburetor would change the horsepower produced by a standard six-cylinder engine employed a 2^4 factorial design listed below. When this design's runs were carried out, the observed independent responses y are as listed below.

Treatment				Response
A	B	C	D	y
−	−	−	−	14.8
+	−	−	−	24.8
−	+	−	−	12.3
+	+	−	−	20.1
−	−	+	−	13.8
+	−	+	−	22.3
−	+	+	−	12.0
+	+	+	−	20.0
−	−	−	+	16.3
+	−	−	+	23.7
−	+	−	+	13.5
+	+	−	+	19.4
−	−	+	+	11.3
+	−	+	+	23.6
−	+	+	+	11.2
+	+	+	+	21.8

   (a) Which dimension is the most critical in influencing the response?
   (b) Assume $\sigma^2 = 4$ and data follow a normal distribution, make a 95% confidence interval statement for the effect of the most important dimension.
   (c) Construct a normal probability plot of the estimated effects and interpret your result.
   (d) By pooling the sum of squares corresponding to nonsignificant effects, obtain an estimate of σ^2. What then is a 95% confidence interval for the effects in part (a)?

2. A chemist conducts an experiment on chemical yield using four factors, each at two levels. The experiment was completely randomized and the factors used were temperature A, reaction time B, catalyst C, and concentration % D. The whole experiment was replicated twice. The results obtained are given below.

A	B	C	D	Treatments	y		A	B	C	D	Treatments	y	
−1	−1	−1	−1	1	57	61	−1	−1	−1	1	d	49	54
1	−1	−1	−1	a	49	55	1	−1	−1	1	ad	53	62
−1	1	−1	−1	b	57	52	−1	1	−1	1	bd	64	59
1	1	−1	−1	ab	46	56	1	1	−1	1	abd	61	55
−1	−1	1	−1	c	50	59	−1	−1	1	1	cd	58	62

A	B	C	D	Treatments	y		A	B	C	D	Treatments	y	
1	−1	1	−1	*ac*	50	53	1	−1	1	1	*acd*	53	59
−1	1	1	−1	*bc*	49	51	−1	1	1	1	*bcd*	57	64
1	1	1	−1	*abc*	52	58	1	1	1	1	*abcd*	60	63

(a) Analyze these data using $\alpha = 0.01$.
(b) Construct a normal probability plot of the residuals and interpret your result.
(c) Plot the residuals versus predicted yield. Interpret this plot.

3. Refer to Problem 2.
   (a) Find the standard error of the estimates of the factor effects.
   (b) Determine 95% confidence intervals for the factor effects.
   (c) Use the confidence intervals in part (b) to determine which effects are significant, and check if your conclusions agree with those in Problem 2.
4. Following are the results from a pilot study of process development.

A	B	C	D	y
−1	−1	−1	−1	68
1	−1	−1	−1	62
−1	1	−1	−1	76
1	1	−1	−1	81
−1	−1	1	−1	77
1	−1	1	−1	66
−1	1	1	−1	79
1	1	1	−1	83
−1	−1	−1	1	64
1	−1	−1	1	73
−1	1	−1	1	75
1	1	−1	1	85
−1	−1	1	1	81
1	−1	1	1	80
−1	1	1	1	89
1	1	1	1	92

   (a) Prepare a partial ANOVA table for these data. How can you complete this ANOVA without adding any new treatments to the above pilot study?
   (b) Construct a normal probability plot of the estimated effects and determine which effects are significant.
   (c) Construct a normal probability plot of the residuals and check the normality assumption.
   (d) Plot the residuals versus predicted yield. Interpret this plot.
5. Refer to Problem 4.
   (a) Pooling the sum of squares corresponding to nonsignificant effects obtained in 4(b), estimate the error variance σ^2.
   (b) Find the standard error of the estimates of the main effects that are significant.

(c) Determine 99% confidence intervals for the main effects that were found to be significant in part (b).

6. Refer to the data in Problem 4 and use the relevant information from Problems 4 and 5 to prepare the revised ANOVA table. Draw your conclusions and then check if they match your conclusions in Problem 5.

18.5 Blocking in the 2^k Factorial Design

It is never possible to conduct experiments in an environment where all sources of variability are eliminated. However, it is often possible to control some sources of variability, thereby establishing environments within which the experimental variability is decreased. In Chapter 17, we studied randomized complete block designs and Latin square designs to reduce experimental variability due to some nuisance variables. This was accomplished by partitioning the design into subsets or *blocks* of experiments within which the experimental environment is held as constant as possible. In this section, we discuss a similar concept for 2^k factorial designs.

18.5.1 Confounding in the 2^k Factorial Design

Whenever the number of factors k becomes large, the number of treatments in a complete replication of a 2^k factorial design becomes so very large that it is usually not possible to run a full replication of the factorial design in one block. For example, we may not have enough raw material from one production batch to run the whole experiment, and there may be significant variation among raw materials from different batches so that it would justify performing the complete experiment in smaller blocks. This makes some interaction effects inseparable from the block effects because we cannot determine which is which. When this happens, we say that the interaction effects are *confounded* with the block effects.

To describe blocking arrangements when using a 2^k factorial, we use notation derived from the runs of a 2^k factorial. To illustrate, consider again a full 2^3 experiment given below in standard order.

Run order	A	B	C	Notation for runs	Observations
1	$-$	$-$	$-$	1	y_1
2	$+$	$-$	$-$	a	y_2
3	$-$	$+$	$-$	b	y_3
4	$+$	$+$	$-$	ab	y_4
5	$-$	$-$	$+$	c	y_5
6	$+$	$-$	$+$	ac	y_6
7	$-$	$+$	$+$	bc	y_7
8	$+$	$+$	$+$	abc	y_8

For example, we will denote run 4 by ab. This notation may be looked at as follows: ab mentions factors A and B but not C, so that the run uses A at its high level (+), B at its high level (+), and C at its low level (−). The notation for the other runs of the 2^3 design follows in similar manner. Finally, we will sometimes use the notation '− a' which simply reverses all signs in the run a. So, for the 2^3 case, we would use $-a$ for the run $(-++)$.

Now, to illustrate blocking, consider a 2^3 factorial experiment performed in two blocks, where the different treatments are assigned to the different blocks as shown below. The experiment is run in a random order.

Block I	Block II
1	a
ab	b
ac	c
bc	abc

In this arrangement we can easily see that the contrast between the two blocks is equal to the three-factor interaction effect ABC. That is, the estimate of the three-factor interaction effect is

$$(\widehat{ABC}) = \frac{1}{4}(-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8)$$

which can be written as

$$(\widehat{ABC}) = \frac{1}{4}(y_2 + y_3 + y_5 + y_8) - \frac{1}{4}(y_1 + y_4 + y_6 + y_7),$$

and this is clearly a contrast between the two blocks. We then say that the three-factor interaction effect ABC and the block effects are *completely confounded.* Also it can easily be checked that all main effects and all the two factor interaction effect contrasts are orthogonal to the block effect contrasts, so *none of the main effects and the two-factor interactions are affected by the block effects,* whereas the three-factor interaction is completely confounded with the block effects.

Note that here we have lost complete information on the three-factor interaction, but the information on other effects is retained and available as usual. It is important to remember that when there are k factors to be analyzed in a 2^k factorial with r replications, the estimated variance of main effects or interaction effects is given by

$$\frac{\hat{\sigma}^2}{r2^{k-2}} = \frac{S^2}{r2^{k-2}} \tag{18.5.1}$$

It is also important to note that the blocks should be created in such a way that only higher-order interactions are confounded. The block with treatment (1) is usually referred to as the *principal block.* The interaction (or interactions in case a replicate is divided into $2^h, h \geq 1$, blocks) that divides the full replicate into blocks is called the generator (or generators). For instance, in the example above, the interaction ABC is a generator. Finally, we may remark here that if the blocks are not constructed appropriately, then more interactions may be confounded with the block effects than desired.

Now to further illustrate the concept of confounding, we consider an example of a 2^3 factorial design in which the three-factor interaction is completely confounded with the block effects.

Example 18.5.1 (A special alloy manufacturing experiment) *A special alloy is prepared to make various parts of jet turbine aircraft engines. In order to avoid cracking in the finished parts, which can cause irreversible engine failure, an experiment using three* *factors is planned. The three factors considered important are (A) pouring temperature, (B) amount of grain refining, and (C) final product treatment. The experiment is performed with two replicates of a 2^3 factorial experiment design. Further, due to some variation, each replication is divided into two blocks. Then, the final product is tested under unusual stress, and the length of crack in hundredth mm is recorded. The experiment was carried out using the runs in random order. The plan of the experiment and the data obtained are shown in Table 18.5.1. Under this plan the three-factor interaction effect is completely confounded with the block effects.*

Table 18.5.1 The plan of the experiment and the data obtained.

Replication I				Replication II			
Block I		Block II		Block III		Block IV	
1	14.16	*a*	13.78	1	13.76	*a*	12.89
ab	10.08	*b*	12.14	*ab*	11.30	*b*	11.93
ac	9.29	*c*	11.97	*ac*	9.37	*c*	12.47
bc	8.05	*abc*	9.70	*bc*	7.80	*abc*	9.39
Total	41.58		47.59		42.23		46.68

Our plan is to do the following:

1. Estimate the main effects and two factor interaction effects.
2. Prepare an ANOVA table for these data and verify if any of the main effects or two-factor interaction effects are significant at the 5% level of significance.
3. Examine if it was necessary to divide each replication into two blocks each.

Solution: To prepare the ANOVA table and estimate the factorial effects, we can use either the table of contrast coefficients for the 2^3 factorial or Yates's algorithm (to be discussed in Section 18.5.2). For now, in this example, we analyze the data using contrast coefficients (see Table 18.3.3). (Later, we will analyze another set of data using Yates's method.) We first find the response total for each treatment combination. That is, $(1) = 27.92$, $(a) = 26.67$, $(b) = 24.07$, $(ab) = 21.38$, $(c) = 24.44$, $(ac) = 18.66$, $(bc) = 15.85$, and $(abc) = 19.09$.

First, the estimates of main effects and interaction effects are

$$\hat{A} = (-27.92 + 26.67 - 24.07 + 21.38 - 24.44 + 18.66 - 15.85 + 19.09)/8 = -0.81$$

$$\hat{B} = (-27.92 - 26.67 + 24.07 + 21.38 - 24.44 - 18.66 + 15.85 + 19.09)/8 = -2.16$$

$$\hat{C} = (-27.92 - 26.67 - 24.07 - 21.38 + 24.44 + 18.66 + 15.85 + 19.09)/8 = -2.75$$

$$\widehat{AB} = (27.92 - 26.67 - 24.07 + 21.38 + 24.44 - 18.66 - 15.85 + 19.09)/8 = 0.95$$

$$\widehat{AC} = (27.92 - 26.67 + 24.07 - 21.38 - 24.44 + 18.66 - 15.85 + 19.09)/8 = 0.175$$

$$\widehat{BC} = (27.92 + 26 : 67 - 24.07 - 21.38 - 24.44 - 18.66 + 15.85 + 19.09)/8 = 0.12$$

Second, from 18.2.5 we have ($r=2$, $k=3$)

$$\text{Sum of squares due to } A = 2(2)(-0.81)^2 = 2.6244$$

$$\text{Sum of squares due to } B = 2(2)(-2.16)^2 = 18.6624$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$\text{Sum of squares due to } BC = 2(2)(0.12)^2 = 0.0576$$

$$\text{Sum of squares due to blocks} = [(41.58)^2 + \cdots + (46.68)^2]/4 - (178.08)^2/16 = 6.99$$

The ANOVA for the data in Table 18.5.1 is as given in Table 18.5.2.

Table 18.5.2 ANOVA table for the data in Table 18.5.2.

Source	*d.f.*	*SS*	*MS*	*F*-ratio	*p*-value
Blocks	3	6.99	2.33	10.33	0.008
Main effects	3	51.54	17.18	79.91	0.0000
2 Factor interaction	3	3.80	1.267	5.89	0.0321
Residual error	6	1.29	0.215		
Total	15	63.62			

Some further analysis

Term	Effect	Regression coefficient	*SE* coefficient	*t*-value	*p*-value
A	−0.81	−0.405	0.1157	−3.49	0.013
B	−2.16	−1.08	0.1157	−9.34	0.000
C	−2.75	−1.37	0.1157	−11.88	0.000
AB	0.95	0.475	0.1157	4.09	0.006
AC	0.175	0.088	0.1157	0.76	0.475
BC	0.13	0.065	0.1159	0.54	0.612

Note that the estimates of the regression coefficients are one-half of the effects because "regression coefficients" measure change when an associated variable is increased by one unit, whereas an "effect" measure change in the response when that variable is increased from −1 to 1 (i.e., change is 2 units). Further, $\hat{V}(\text{coeff}) = \hat{V}((1/2) \text{ effect}) \Rightarrow SE(\text{coeff}) = SE((1/2) \text{ effect})$, and so on.

Now from the p-values listed in the second half of Table 18.5.2, we see that all main effects and the two-factor interaction AB are highly significant.

Third, from Table 18.5.2, we note the p-value corresponding to the blocks is 0.008, which is less than 0.05. Thus, we reject the null hypothesis of no block effects at significance level 0.05. That is, we conclude that the blocking in this experiment was necessary.

Example 18.5.2 (Using MINITAB and R) *This example is concerned with a replicated 2^3 factorial in four blocks of four treatments each. A study was performed to determine the effects of texturing on the breaking strength of an artificial fiber. The process variables selected for study were X_1 (spindle speed), X_2 (temperature of plates), and X_3 (amount of* *twist). A 2^3 factorial design in* r = *2 replicates was chosen. Since only four experiments could be run during a single day, each 2^3 factorial design was partitioned into two blocks of four runs each, and the program completed on four separate days. The experiments were randomly run within each day. The 2^3 design was partitioned as illustrated in Table 18.5.3; that is, the contrast of the three-factor interaction effect was used to block the design. The results are also displayed in Table 18.5.3 (blocks represent different days). We wish to analyze this set of data using MINITAB and R.*

Table 18.5.3 Two replications (each replication in two blocks) of a 2^3 experiment.

Block I		Block II		Block III		Block IV	
1	18.8	*a*	19.8	1	17.7	*a*	18.4
ab	18.0	*b*	11.8	*ab*	15.0	*b*	12.8
ac	19.0	*c*	22.7	*ac*	19.8	*c*	23.8
bc	20.6	*abc*	13.6	*bc*	19.5	*abc*	14.8
Total	76.4		67.9		72.0		69.8

MINITAB

To analyze the data of a 2^k factorial design using MINITAB, we first need to generate the design that will confound the minimum number of interactions. We proceed as follows:

1. From the Menu bar select **Stat** > **DOE** > **Factorial** > **Create Factorial Design . . .** .
2. In the dialog box that appears select **2-level factorial (default generators)**. Next, select appropriate number of factors and click **Designs**.
3. Since we are using complete replications, select full factorial design, enter zero for **Number of center points per block**, select 2 (i.e., the number of replications) for **Number of replicates for corner points**, and enter 4 for **Number of blocks** (i.e., the total number of blocks in the whole experiment), and click **OK**.
4. In the dialog box to "Create Factorial Design," select options, uncheck **Randomized runs** (if it is checked), and click **OK**. Then, select factors to select the code of low and high level and the name of the factors (by default low and high level in MINITAB as −1 and 1, respectively, which you may leave alone since they do not affect the analysis, (note that in this case the regression coefficients are equal to half the estimates of the effects, since effects represent the change as the level changes 2 units from −1 to 1, whereas regression coefficients are the rate of change per unit), and click **OK**. Again click **OK**. The design generated by MINITAB appears in the **Worksheet** window. The order of the treatments will be exactly the same as in Table 18.5.3.
5. Enter the data in the next available column and again, from the Menu bar select **Stat** > **DOE** > **Factorial** > **Analyze Factorial Design . . .** . Another dialog box "Analyze Factorial Design" appears where you may select any options that you would like to see in your analysis output. For example, under **Terms . . .** , select in the right box those main effects and interactions that you would like to estimate. Then click **OK**. The final analysis output appears in the session window as shown here.

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Model	9	171.001	96.63%	171.001	19.0001	19.12	0.001
Blocks	3	10.027	5.67%	10.027	3.3423	3.36	0.096
Linear	3	106.122	59.97%	106.122	35.3740	35.59	0.000
A	1	5.406	3.05%	5.406	5.4056	5.44	0.058
B	1	71.826	40.59%	71.826	71.8256	72.26	0.000
C	1	28.891	16.33%	28.891	28.8906	29.07	0.002
2-Way Interactions	3	54.852	31.00%	54.852	18.2840	18.40	0.002
A*B	1	0.456	0.26%	0.456	0.4556	0.46	0.524
A*C	1	54.391	30.74%	54.391	54.3906	54.72	0.000
B*C	1	0.006	0.00%	0.006	0.0056	0.01	0.942
Error	6	5.964	3.37%	5.964	0.9940		
Total	15	176.964	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.996975	96.63%	91.57%	42.4089	76.04%

Coded Coefficients

Term	Effect	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant		17.881	0.249	(17.271, 18.491)	71.74	0.000	
Blocks							
1		1.219	0.432	(0.162, 2.275)	2.82	0.030	1.50
2		–0.906	0.432	(–1.963, 0.150)	–2.10	0.081	1.50
3		0.119	0.432	(–0.938, 1.175)	0.28	0.792	1.50
A	–1.163	–0.581	0.249	(–1.191, 0.029)	–2.33	0.058	1.00
B	–4.238	–2.119	0.249	(–2.729, –1.509)	–8.50	0.000	1.00
C	2.688	1.344	0.249	(0.734, 1.954)	5.39	0.002	1.00
A*B	0.337	0.169	0.249	(–0.441, 0.779)	0.68	0.524	1.00
A*C	–3.687	–1.844	0.249	(–2.454, –1.234)	–7.40	0.000	1.00
B*C	0.038	0.019	0.249	(–0.591, 0.629)	0.08	0.942	1.00

The residual analysis graph in Figure 18.5.1 does not indicate any unusual violation of the model assumptions. The estimated effects and their p-values obtained by employing the t-test indicates that the main effects B and C are highly significant. The main effect A is not significant at the 5% level of significance, but it is significant at 10% level of significance. The only interaction that is highly significant is AC. R-Sq and R-Sq (adj) are quite high and, combined with p-values, indicate the fitted model is quite adequate. R-Sq (pred.) shows that the prediction capability of the model is 76%.

Further, from the ANOVA table, we see that the estimate of the experimental error variance is $S^2 = 0.9940$ with six degrees of freedom. The 95% confidence limits for the

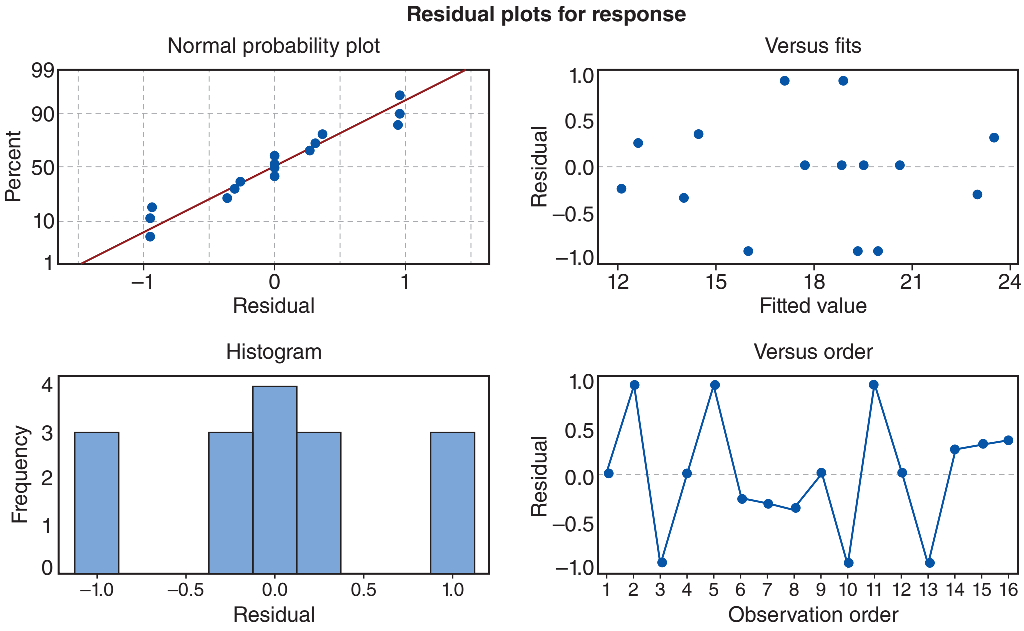


Figure 18.5.1 MINITAB printout of the residual plots for the data in Table 18.5.3.

factorial effects (see Equation 18.3.8) are given by

$$\{\text{Estimate of effect}\} \pm 2.447\sqrt{\left(\frac{0.9940}{2(4)}\right)} = \{\text{Estimate of effect}\} \pm 0.86$$

Solution:

USING R

The R functions ‘conf.design()’ and ‘aov()’ can be used to generate and run the analysis of a 2^k factorial design as shown in the following R-code.

```
install.packages(“conf.design”); library(“conf.design”)
Response = c(18.8,18.0,19.0,20.6,19.8,11.8,22.7,13.6,17.7,15.0,19.8,
19.5,18.4,12.8,23.8,14.8)

#Generate 2^3 design with 4 blocks. ‘G’ indicates the aliasing structure, Blocks =
A*B*C (see Section 18.6) and ‘p’ indicates the number of levels in each factor
Design1 = conf.design(G = c(1,1,1), p = 2, block.name = “Blocks”,
treatment.names = c(“A”,“B”,“C”))

#Replicate above design to get Block III and Block IV
Design2 = conf.design(G = rbind(c(1,1,1), c(1,1,1)), p = 2,
block.name = “Blocks”, treatment.names = c(“A”,“B”,“C”))
```

```
#Final design with Response data
Design = rbind(Design1,Design2)
Data = cbind(Response,Design)


#Fitting the model
model = aov(Response ~ Blocks+A*B*C, data=Data)
model
anova(model)


#R output: Analysis of Variance Table
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Blocks	3	10.03	3.34	3.36	0.0962
A	1	5.41	5.41	5.44	0.0585
B	1	71.83	71.83	72.26	0.0001
C	1	28.89	28.89	29.07	0.0017
A:B	1	0.46	0.46	0.46	0.5236
A:C	1	54.39	54.39	54.72	0.0003
B:C	1	0.01	0.01	0.01	0.9425
Residuals	6	5.96	0.99		

The above R output conveys the same information we observed in MINITAB (ANOVA) results and therefore the conclusions stay the same.

We now take a very brief look at designs where we would conduct a 2^k factorial experimental design in 2^{k-1} blocks of two runs each. The 2^{k-1} blocks, for example, might represent 2^{k-1} different operators or 2^{k-1} different batches of raw material. In this case, there will be $(2^{k-1} - 1)$ degrees of freedom assignable to the blocks, leaving $2^{k-1} = [(2^k - 1) - (2^{k-1} - 1)]$ degrees of freedom for estimating factorial effects. These designs are of considerable interest. The 2^3 and 2^4 factorial designs, partitioned into blocks of two treatments each, are displayed in Table 18.5.4. These designs are structured so that the main effects estimates are not affected at all (i.e. not confounded) by differences between the blocks. These designs are sometimes termed *main effect clear* designs. The reader will note that the pairs of treatments comprising the blocks are "complementary" or "fold-over" pairs. The second row in any block has elements that are $(-1) \times$ (elements in the first row of the block). Thus, differences between treatments *within* a block are unaffected by the block means. The 2^{k-1} differences determined from within the 2^{k-1} blocks supply all the necessary information for estimating the k main effects, clear of block effects. For more details of these of designs, the reader is referred to Box et al. (1978).

Table 18.5.4 Runs of the 2^3 and 2^4 factorial design in blocks of two treatments each.

		X_1	X_2	X_3
Block	(i)	−	−	−
		+	+	+
	(ii)	+	−	−
		−	+	+
	(iii)	−	+	−
		+	−	+
	(iv)	+	+	−
		−	−	+

		X_1	X_2	X_3	X_4
Block	(i)	−	−	−	−
		+	+	+	+
	(ii)	+	−	−	−
		−	+	+	+
	(iii)	−	+	−	−
		+	−	+	+
	(iv)	+	+	−	−
		−	−	+	+
	(v)	−	−	+	−
		+	+		+
	(vi)	+	−	+	−
		−	+	−	+
	(vii)	−	+	+	−
		+	−	−	+
	(viii)	+	+	+	−
		−	−	−	+

18.5.2 Yates's Algorithm for the 2^k Factorial Designs

Yates's algorithm is a method that speeds up the computations required in a 2^k factorial or fractional factorial design, particularly when one is interested in estimating all or most of the effects. If, however, one is interested in estimating only a few effects and k is large, then it may be easier to use the table of contrast coefficients (see Table 18.3.3). To employ the algorithm, it is essential that the 2^k factorial design be written down in standard order with the first column of the design array A_1 consisting of alternating minus and plus signs, the second column A_2 with alternating pairs of minus and plus signs, the third column A_3 with alternating groups of four minus and plus signs, and so on until the kth column A_k contains 2^{k-1} minus signs followed by the same number of plus signs. A minus sign is the first element in each column. The 2^4 factorial in standard or Yates's order is shown in Table 18.4.2. The observation y_i recorded for each of the $i = 1, 2, \ldots, 2^k$ treatments (or in the case of more than one replicates the total T_i of the observations for each treatment) is recorded in a column following the design matrix. The observations are now grouped into separate pairs and by a series, first of additions of observations in pairs and then of subtractions (always subtracting the first entry from the second) of observations in pairs. Thus, a first column of data is constructed with the 2^k entries. This new column is in turn used to construct a second column in the same manner, and the *procedure is repeated until a total of k new columns have been formed.* Specifically, one proceeds as follows:

1. The pairs of entries are algebraically summed, the sum of each successive pair providing successively the first 2^{k-1} entries in the new column.

2. The pairs of entries are successively algebraically summed after changing the sign of the first entry in each of the pairs. These new successive sums provide the final 2^{k-1} entries in the new column.
3. The top entry in the *k*th column will be the total of all the observations. Dividing by the total number of observations $N = r2^k$, where r is the common number of replicates of the 2^k treatments, allows for the calculation of the grand average $\bar{Y}$.
4. The estimated effects are given by the remaining $2^k - 1$ entries in the final column of the algorithm divided by $N/2 = r2^{k-1}$.
5. The estimated effects are identified by noting the "plus signs" in the treatment identification on the same line in the design matrix.
6. The individual degree of freedom sum of squares for each of the $2^k - 1$ estimated effect is given by squaring the entries in the last column and dividing by $N = r2^k$. Note that the first element, squared and divided by N, is the correction factor.

To check the computations, let $W = \sum T^2$ be the sum of squares of the treatment totals. Then, the sum of squares of the entries in the *k*th column should be equal to $r2^k W$.

PRACTICE PROBLEMS FOR SECTION 18.5

1. Refer to Problem 2 of Section 18.4. (a) Construct a design with two blocks of eight observations each using the interaction $ABCD$ as a design generator. (b) Analyze the data in the principal block with two replications.
2. Refer to Problem 1 above and Problem 2 of Section 18.4. Suppose that the two halves of the first replication were run in two separate labs. Analyze the data assuming that the labs are considered as blocks. (*Hint*: interaction $ABCD$ is confounded with the block effects)
3. Consider a 2^5 factorial design. Construct a design with four blocks of eight observations each with $ABCD$ and BCE confounded, and state which of the other interactions are confounded with blocks.
4. Suppose that an experiment is run with five factors A, B, C, D, and E, where each factor is at two levels, and that the resulting data are as shown below:

(1) $=13$	$a=17$	$b=23$	$ab=34$	$c=29$	$ac=21$	$bc=32$	$abc=36$
$d=25$	$ad=32$	$bd=39$	$abd=29$	$cd=31$	$acd=44$	$bcd=37$	$abcd=28$
$e=33$	$ae=38$	$be=29$	$abe=22$	$ce=35$	$ace=27$	$bce=43$	$abce=29$
$de=42$	$ade=44$	$bde=41$	$abde=47$	$cde=35$	$acde=46$	$bcde=49$	$abcde=53$

Suppose that this experiment was run in four blocks with $ABCD$ and BCE confounded. Analyze these data using $\alpha = 0.05$ (see Problem 3).

5. Use the data of Problem 1 of Section 18.4. Suppose that the experiment was run in two blocks of eight observations each, with ABC confounded with blocks. State the design and analyze the data of this experiment.
6. Construct a 2^6 design in four blocks in such a way that two three-factor and one four-factor interactions are confounded with blocks.
7. Analyze the data in Problem 1 of Section 18.4 using Yates's method.

18.6 The 2^k Fractional Factorial Designs

As we stated earlier, whenever the number of factors k becomes large, the number of treatments required by the 2^k factorial designs becomes burdensomely large. For example, a complete replicate of the 2^7 design requires 128 treatment runs. When completed, the data from a 2^k design provides, in addition to estimates of the k main effects and the $k(k-1)/2$ two-factor interactions, estimates of all the three-factor and higher-order effects. However, it is often the case that the three-factor and higher-order effects can be assumed *a priori* to be zero, or at least to be small relative to the lower-order effects. When this is true, only a fraction of the 2^k design need be employed. Here our discussion will be restricted to the one-half and one-quarter replicate designs, or the so-called 2^{k-1} and 2^{k-2} fractional factorial designs, respectively. Discussion of the general 2^{k-h} fractional factorial design, called the $1/2^h$ fraction of the 2^k factorial design, is beyond the scope of this book, but we refer the interested reader to Box and Hunter (1961), Box et al. (1978), and Montgomery (2009a,b).

18.6.1 One-half Replicate of a 2^k Factorial Design

The 2^{k-1} fractional factorial designs may be constructed by first partitioning the 2^k factorial into two blocks of 2^{k-1} runs each, using the highest order interaction contrast as the generator. Each block is then a 2^{k-1} design. For example, to construct the 2^{4-1} design, one begins (as demonstrated in Table 18.6.1) by writing down the full 2^4 factorial design in standard order and then partitioning the design into two blocks of eight runs each. In this case the ($ABCD$) four-factor interaction contrast vector is used as the generator, also commonly called a *defining contrast*. The 2^{4-1} design, *design* 1 in Table 18.6.1, *consists of the eight runs of the 2^4 factorial* that contain a plus sign in the four-factor interaction contrast vector $ABCD$, while *design* 2 consists of the eight runs possessing a minus sign in this vector. Since we are using the $ABCD$ column in this way, the *generators* of these fractional factorials are $+ABCD$ and $-\ ABCD$, respectively.

With only 2^{k-1} treatments in either design 1 or design 2, it is obviously impossible to estimate all the (2^k-1) individual effects in the factorial model. However, 2^{k-1} orthogonal contrasts can be calculated, and it is important to identify the confounded factorial effects estimated by these statistics. The confounding pattern, or alias structure, is best explained by an example. In Table 18.6.2, we see the 2^{4-1} fractional factorial design with *generator* $+ABCD$ along with a set of corresponding observations. The design has been listed in standard Yates's order with respect to variables A, B, and C. The design was run in random order.

To determine some estimates, we first note (see Table 18.6.2) that $k=4$ and that we are using a one-half replication of the 2^4 design, so that any effect estimator is a combination of $(1/2)\times 2^4=8$ observations, which can be expressed as $\bar{Y}_+ - \bar{Y}_-$, as discussed previously, where $\bar{Y}_+$ and $\bar{Y}_-$ are means of four observations with factor A at higher and lower levels, respectively. For example, we can express $\hat{A}=\bar{Y}_+ - \bar{Y}_-$ as

$$\hat{A}=\frac{1}{4}(\{A\}\times y)=\frac{1}{4}(-9.4+16.7-\cdots-4.1+11.3)=4.70$$

Here $\{A\}$ represents the contrast coefficients associated with the main effect of the factor A, and the data are as in the y-column of Table 18.6.2. To estimate the BCD effect,

Table 18.6.1 The full 2^4 factorial design and two 2^{4-1} designs obtained by the generators $+ABCD$ and $-ABCD$.

2^4 Factorial					2^{4-1}(Design 1)				2^{4-1} (Design 2)			
A	B	C	D	$ABCD$	A	B	C	D	A	B	C	D
−	−	−	−	+	−	−	−	−	+	−	−	−
+	−	−	−	−	+	+	−	−	−	+	−	−
−	+	−	−	−	+	−	+	−	−	−	+	−
+	+	−	−	+	−	+	+	−	+	+	+	−
−	−	+	−	−	+	−	−	+	−	−	−	+
+	−	+	−	+	−	+	−	+	+	+	−	+
−	+	+	−	+	−	−	+	+	+	−	+	+
+	+	+	−	−	+	+	+	+	−	+	+	+
−	−	−	+	−	Generator: $+\,ABCD$				Generator:$-\ ABCD$			
+	−	−	+	+	defining				defining			
−	+	−	+	+	relation $I+1234$				relation $I-1234$			
+	+	−	+	−								
−	−	+	+	+								
+	−	+	+	−								
−	+	+	+	−								
+	+	+	+	+								

Table 18.6.2 A 2^{4-1} fractional factorial design.

Design				Observations	Some factorial effect contrast coefficients			
A	B	C	D	y	A	BCD	AB	CD
−	−	−	−	9.4	−	−	+	+
+	−	−	+	16.7	+	+	−	−
−	+	−	+	12.6	−	−	−	−
+	+	−	−	15.4	+	+	+	+
−	−	+	+	5.2	−	−	+	+
+	−	+	−	6.7	+	+	−	−
−	+	+	−	4.1	−	−	−	−
+	+	+	+	11.3	+	+	+	+

we find that the contrast coefficients BCD are identical to those already used to estimate the A effect, and we denote this by $A = BCD$. Hence, we say that the A and BCD effects are confounded. In fact, it can be shown that $E[\frac{1}{4}(\sum\{A\} \times y)] = A + BCD$. Similarly, the contrast coefficients for the AB interaction effect are identical to those for the CD interaction effect, so AB and CD interaction effects are confounded. In this example, the four main effects are indeed each confounded with a single three-factor interaction, and that the six two-factor interactions are confounded in pairs.

A convenient method for determining the confounding pattern for 2^{k-1} fractional factorials, and hence, the expected value of the $2^{k-1}-1$ orthogonal factorial contrasts, is provided by the design's *defining relation.* In the example above, the design generator is $+ABCD$ or the "word" $+ABCD$. The defining relation for the designs is then the "sentence" $\boldsymbol{I}+\boldsymbol{ABCD}$, where the symbol $\boldsymbol{I}$ is called the identity (i.e. the coefficients of $\boldsymbol{I}$ are all pluses) and the sentence consists of two words. In general, for the 2^{k-h} fractional factorials, there will be h generators and the defining relation will be a sentence containing 2^h words. Multiplying the sentence of defining relation by, say $\boldsymbol{A}$, gives $\boldsymbol{I}\times\boldsymbol{A}+\boldsymbol{A}^2\boldsymbol{BCD}$.

We now adopt the rule that *any* symbol that appears to an even power converts to the identity $\boldsymbol{I}$ because, if we square the contrast coefficients of any effect, we always get all plus signs. Thus, dropping the identity $\boldsymbol{I}$, we have $\boldsymbol{I}\times\boldsymbol{A}+\boldsymbol{A}^2\boldsymbol{BCD}=\boldsymbol{A}+\boldsymbol{BCD}$. Similarly multiplying the defining relation by $\boldsymbol{AB}$ gives $\boldsymbol{AB}+\boldsymbol{CD}$. The confounded effects are usually called the *aliases*, so A, BCD are aliases and AB, CD are aliases. Thus, in this example, we have a set of eight groups of aliases: $\{I, ABCD; A, BCD; B, ACD; C, ABD; D, ABC; AB, CD; AC, BD; AD, BC\}$. If the generator of the design had been $-\ ABCD$, then the defining relation would be $\boldsymbol{I}-\ \boldsymbol{ABCD}$. For the full 2^k factorial designs, the defining relation is simply $\boldsymbol{I}$.

The two one-half replicates of a 2^4 design generated by the defining relation $\boldsymbol{I}+\boldsymbol{ABCD}$ and $\boldsymbol{I}-\ \boldsymbol{ABCD}$ are the two designs

$$\{1,\ ab, ac, ad, bc, bd, cd, abcd\} \text{ and } \{-a, -b, -c, -d, -abc, -abd, -acd, -bcd\}$$

where the first one-half replicate containing the treatment (1) is usually called the principal block (see Table 18.6.1).

The analysis of a 2^{k-h} fractional factorial is accomplished by initially considering the data as having been provided by a 2^p full factorial design, where $p=k-h$ is some convenient subset of the k factors. The (2^p-1) factorial effects are then estimated using Yates's algorithm or a table of contrast coefficients. Using only the p factors, we then label each factorial effect with its *naive* name or word (by ignoring h letters from the word representing an effect). The expected value of each estimate is determined by multiplying the defining relations of the design, respectively, by its *naive* word and restoring the ignored letters. The assumption is usually made that three-factor and higher-order effects may be ignored. This assumption simplifies the confounding pattern, also known as alias structure.

We illustrate this alias structure by considering a simple example of the 1/2 replicate of a 2^4 design, namely the design that uses the following runs:

$$1, ab, ac, ad, bc, bd, cd, abcd$$

Now, the naïve words are obtained by ignoring the letter (d), say, then writing the halved replicate after some rearrangements as

$$1, a(d), b(d), c(d), ab, ac, bc, abc(d)$$

where the ignored letter is written in parentheses. This clearly represents a full replicate in factors A, B, and C. Then, after calculating the effects and their sums of squares in the same manner as for a 2^3 factorial experiment, the ignored letter is reintroduced by using the alias structure. We further illustrate this concept with a numerical example.

Table 18.6.3 Estimates of effects of the 2^{4-1} fractional factorial design.

Treatments	Observations	(1)	(2)	(3)	Effect estimates (3)/4	Sum of squares $(3)^2/8$	Aliases
1	9.4	26.1	54.1	81.4	10.175		$I+ABCD$
$a(d)$	16.7	28.0	27.3	18.8	4.7	44.18	$A+BCD$
$b(d)$	12.6	11.9	10.1	5.4	1.35	3.645	$B+ACD$
ab	15.4	15.4	8.7	1.2	0.3	0.18	$AB+CD$
$c(d)$	5.2	7.3	1.9	−26.8	−6.7	89.78	$C+ABD$
ac	6.7	2.8	3.5	−1.4	−0.35	0.245	$AC+BD$
bc	4.1	1.5	−4.5	1.6	0.4	0.32	$BC+AD$
$abc(d)$	11.3	7.2	5.7	10.2	2.55	13.005	$ABC+D$

Example 18.6.1 (Analyzing a half replication of a 2^4 design using Yates's algorithm) *We analyze the data in Table 18.6.2 of the replicate of a 2^4 design.*

Solution: From Table 18.6.2, we have the following data:

1	ab	ac	ad	bc	bd	cd	$abcd$
9.4	15.4	6.7	16.7	4.1	12.6	5.2	11.3

Now ignoring the letter d and applying Yates's algorithm for 2^3 factorial design, we obtain Table 18.6.3.

Note that the first entry only has divisor 8, since the first entry in column 3 is not a contrast, but the sum of all eight observations. Now assuming that three-factor interactions are negligible, the estimates of all main effects are given by (see the aliases in Table 18.6.3)

$$\widehat{A}=4.7,\ \widehat{B}=1.35,\ \widehat{C}=-6.7,\ \widehat{D}=2.55$$

If we now assume that all two-factor interactions are negligible, then we can perform the testing of hypothesis about the main effects by using as the error sum of squares (with three degrees of freedom) the total of the sum of squares of all the two-factor interactions, which is $(0.18+0.245+0.32)=0.745$. Thus the error mean square is 0.248. The mean square errors associated with main effects A, B, C, and D, each with one degree of freedom, are 44.18, 3.645, 89.78, and 13.005, respectively. Hence, using the appropriate F-statistic, which under the null hypothesis that an effect is zero has the $F_{1,3}$-distribution, we can easily show that all main effects are highly significant.

Finally, from Table 18.6.3 of the half replicate of a 2^4 design, we note that from the last entry in the Aliases column that $\boldsymbol{D}$ and $\boldsymbol{ABC}$ are aliases. Thus, we can construct the one-half replicate of a 2^4 design simply by writing a full 2^3 design for factors A, B, and C and then adding the column of signs associated with the ABC interaction and labeling it as factor D. In general, we can construct the one-half replicate of a 2^k design by writing a full design for 2^{k-1}, then adding the column of signs for the $A_1A_2\cdots A_{k-1}$ interaction and labeling it as factor A_k.

Example 18.6.2 (Analyzing a half replication of a 2^5 design using Yates's algorithm) *A development laboratory is attempting to improve the performance of a packaging machine.*

Five components, say A, B, C, D, E (each a small metal arm of unique shape), have been redesigned, and the objective of the experiments is to determine whether changing one or more of the components will have a salutary effect on the response, the crease retention of the packaging paper. A half replicate of the 2^5 *factorial design was employed, using the generator –ABCDE. The 16 runs were performed in a random sequence. The data, arrayed for our convenience in Yates's order, on variables A, B, C, and D are displayed in Table 18.6.4. (The minus and plus signs are used to denote the standard component and redesigned component of the metal arms.)*

Fifteen orthogonal factorial contrasts can now be estimated from the 16 observations. This is quickly accomplished using Yates's algorithm on four of the five factors, as illustrated for factors *A, B, C,* and *D* in Table 18.6.4. The initial or naive identification for the contrasts is also listed. The expected value for each estimate is determined from the design-defining relation $\boldsymbol{I} - \boldsymbol{ABCDE}$. Assuming now that three-factor and higher-order interaction effects are zero, we obtain orthogonal estimates of the $k = 5$ main effects and of each of the $5(4)/2 = 10$ two-factor interaction effects. Now, when using the defining relation, care must be taken in affixing the proper sign to the estimated effects. For example, the estimated main effect of E is 0.0725 and the estimated AE interaction effect is -0.0425. The reader can check that the contrasts estimating these effects are

$$
\begin{array}{ll}
E \text{ effect:} & (-4.01 + 3.09 + 3.23 + \cdots + 4.47 + 4.44 - 5.26)\ /\ 8 = 0.0725 \\
AE \text{ effect:} & (4.01 + 3.09 - 3.23 - \cdots + 4.47 - 4.44 - 5.26)\ /\ 8 = -0.0425
\end{array}
$$

As another check, we note that from the Aliases column of Table 18.6.4 that

$$(\widehat{BCD} - \widehat{AE}) = 0.0425 \quad \text{and} \quad \widehat{ABCD} - \hat{E} = -0.0725$$

But third and higher-order interactions are assumed to be zero, so the above can be written as

$$-\widehat{AE} = 0.0425, \text{ and } -\hat{E} = -0.0725, \text{ or } \widehat{AE} = -0.0425, \ \hat{E} = 0.0725$$

For a check on arithmetic, recall from the discussion of Yates's algorithm, as applied to this set of data, that

$$16 \sum_{i=1}^{16} y_i^2 = \text{Sums of squares of entries in column 4 (of the Yates algorithm section)} \tag{18.6.1}$$

(see the observations and the entries in column (4) of the Yates's algorithm part of Table 18.6.4). Now the left-hand side 18.6.1 is

$$\text{LHS} = 16 \times \{4.01^2 + 3.09^2 + \cdots + 4.44^2 + 5.26^2\} = 16 \times \{269.0512\} = 4304.8129$$

The right-hand side of 18.6.1 is

$$\text{RHS} = (64.56)^2 + 0^2 + (-1.02)^2 + \cdots + (.34)^2 + (-.58)^2 = 4304.8192$$

that is, LHS = RHS, so that the check 18.6.1 holds.

Table 18.6.4 2^{5-1} fractional factorial design generator $-$ *ABCDE* with defining relation *I*=$-ABCDE$.

						Yates's algorithm						
A	*B*	*C*	*D*	*E*	*y*	(1)	(2)	(3)	(4)	Naive estimates		Aliases
−	−	−	−	−	4.01	7.10	13.88	31.33	64.56	$\bar{y}$ =	4.0350	
+	−	−	−	+	3.09	6.78	17.45	33.23	Zero	$\hat{A}$ =	Zero	*A–BCDE*
−	+	−	−	+	3.23	9.10	13.81	−0.25	−1.02	$\hat{B}$ =	−0.1275	*B–ACDE*
+	+	−	−	−	3.55	8.35	19.42	0.25	6.26	$\widehat{AB}$ =	0.7825	*AB–CDE*
−	−	+	−	+	4.93	6.87	−0.60	−1.07	9.18	$\hat{C}$ =	1.1475	*C*–AB*DE*
+	−	+	−	−	4.17	6.94	0.35	0.05	0.78	$\widehat{AC}$ =	0.0975	*AC–BDE*
−	+	+	−	−	3.62	9.72	0.21	3.11	−0.52	$\widehat{BC}$ =	−0.0650	*BC–ADE*
+	+	+	−	+	4.73	9.70	0.04	3.15	0.68	$\widehat{ABC}$ =	0.0850	*ABC–DE*
−	−	−	+	+	3.77	−0.92	−0.32	3.57	1.90	$\hat{D}$ =	0.2375	*D–ABCE*
+	−	−	+	−	3.10	0.32	−0.75	5.61	0.50	$\widehat{AD}$ =	0.0625	*AD–BCE*
−	+	−	+	−	3.03	−0.76	0.07	0.95	1.12	$\hat{B}D$ =	0.1400	*BD–ACE*
+	+	−	+	+	3.91	1.11	−0.02	−0.17	0.04	$\widehat{ABD}$ =	0.0050	*ABD–CE*
−	−	+	+	−	5.25	−0.67	1.24	−0.43	2.04	$\widehat{CD}$ =	0.2550	*CD–ABE*
+	−	+	+	+	4.47	0.88	1.87	−0.09	−1.12	$\widehat{ACD}$ =	−0.1400	*ACD–BE*
−	+	+	+	+	4.44	−0.78	1.55	0.63	0.34	$\widehat{BCD}$ =	0.0425	*BCD–AE*
+	+	+	+	−	5.26	0.82	1.60	0.05	−0.58	$\widehat{ABCD}$ =	−0.0725	*ABCD–E*

To perform a test of normal plot of estimates, we order the naive estimates shown in Table 18.6.5.

The normal plot of the estimates of the effects in Figure 18.6.1 suggests that only the C and AB effects (labeled as 3 and 12, etc.) are significant. We assume now that only factors 1, 2, and 3 have effects on the response, so that the 2^{5-1} fractional factorial design becomes a replicated 2^3 design in these variables.

Note the usefulness of the fractional factorial designs for screening. Before the experiment, $k = 5$ candidate variables were thought possibly to have a large effect. The experimental strategy located a particular subset of three variables. There were then $\binom{5}{3} = 10$ possible subsets of three. The 2^{5-1} design became a 2^3 factorial, replicated, in any selected subset of three variables. Likewise, if four of the original variables had been found to have important effects, the 2^{5-1} design would have become a 2^4 factorial. There are, of course, five such possibilities. If two of the variables had proved significant, the design would have collapsed into one of the $\binom{5}{2} = 10$ possible 2^2 factorials, replicated four times, and so on. Finally, where only one variable is important, each of the five main effects would become separately estimable.

18.6.2 One-quarter Replicate of a 2^k Factorial Design

For relatively large number of factors, smaller fractions of the 2^k design are quite desirable.

In this section, we consider a one-quarter replicate of a 2^k design, which calls for 2^{k-2} treatment combinations and is usually known as a 2^{k-2} fractional factorial design.

Table 18.6.5 Estimates and their ranks.

Identification of effects	Estimates	Rank order i	$(i-0.5)/15$
C–$ABDE$	1.1475	15	0.9667
AB–CDE	0.7825	14	0.9000
CD–ABE	0.2550	13	0.8333
D–$ABCE$	0.2375	12	0.7667
BD–ACE	0.1400	11	0.7000
AC–BDE	0.0975	10	0.6333
ABC–DE	0.0850	9	0.5667
AD–BCE	0.0625	8	0.5000
BCD–AE	0.0425	7	0.4333
ABD–CE	0.0050	6	0.3667
A–$BCDE$	Zero	5	0.3000
BC–ACE	−0.0650	4	0.2333
$ABCD$–E	−0.0725	3	0.1667
B–$ACDE$	−0.1275	2	0.1000
ACD–BE	−0.1400	1	0.0333

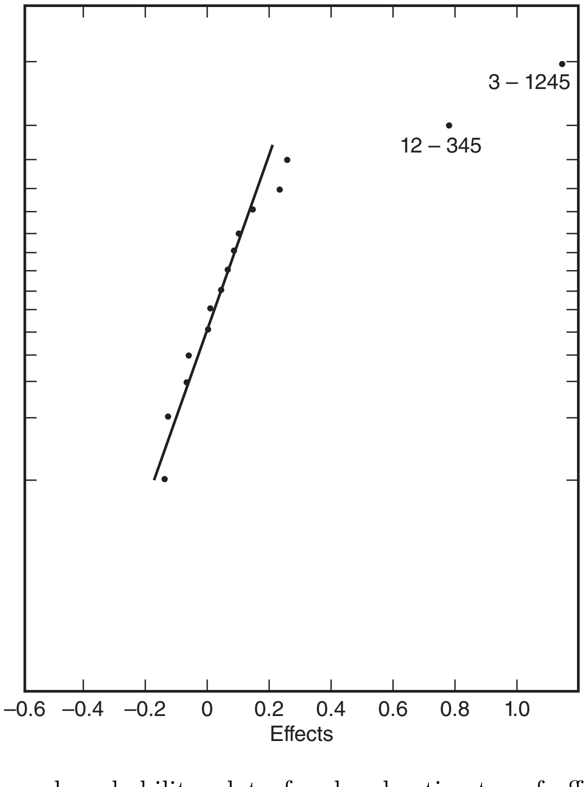


Figure 18.6.1 Normal probability plot of ordered estimates of effects (Table 18.6.4). $A=1$, $B=2$, $C=3$, $D=4$, and $E=5$.

The 2^{k-2} fractional factorial design is constructed by first writing down the 2^{k-2} design for factors $A_1A_2, \ldots,$ and A_{k-2}, then adding two columns of signs associated with two alias and labeling them as factors A_{k-1} and A_k. The two generators, $\pm G_1$ and $\pm G_2$, required for constructing a one-quarter replicate of a 2^k design should be chosen carefully so that the important effects are not confounded with each other, that is, that they are not in the same alias group. This may occur because the defining relation words also contain the generalized interaction $G_1 \times G_2$ that may be an important effect. We illustrate the construction and analysis of a 2^{k-2} fractional factorial design with an example of a 2^{5-2} fractional factorial design.

As previously mentioned, the construction of a 2^{5-2} fractional factorial design is dependent on the selection of two design generators in such a way that the generators and their generalized interaction are of least interest to the experimenter (usually higher-order interactions). If we consider ***ABCD*** and ***ABCDE*** to be two design generators for a 2^{5-2} design, then the sentence for the complete defining relation is $\boldsymbol{I + ABCD + ABCDE + ABCD{*}ABCDE}$, that is, $\boldsymbol{I + ABCD + ABCDE + E}$. This means that we are losing complete information not only for interaction effects ***ABCD*** and ***ABCDE*** but also for the main effect ***E***. Thus, selecting ***ABCD*** and ***ABCDE*** as the two design generators would not be a good choice. Hence, in this case, selecting two generators of three-factor interactions, say ***ABE*** and ***CDE***, having only one letter ***E*** in common may be a better choice because the generalized interaction of the words *ABE* and *CDE* is $\boldsymbol{ABE \times CDE = ABCD}$, a four-factor interaction. Also, in this case, the defining relation is $\boldsymbol{I + ABE + CDE + ABCD}$, and the alias structure for this 2^{5-2} fractional factorial design is as in Table 18.6.6.

Table 18.6.6 Alias structure for the 2^{5-2} fractional factorial design (generators *ABE* and *CDE*).

$A + BE + ACDE + BCD$	$AD + BDE + ACE + BC$
$B + AE + BCDE + ACD$	$AC + BCE + ADE + BD$
$C + ABCE + DE + ABD$	$I + ABE + CDE + ABCD$
$D + ABDE + CE + ABC$	
$E + AB + CD + ABCDE$	

Now, to construct a 2^{5-2} fractional factorial design, we first write down a full 2^3 factorial design for factors A, B, and C. Then, the two factors D and E are added with their associated two columns namely as $D = ABC$ and $E = AB$. The result is the 2^{5-2} design in Table 18.6.7 for the factors *A, B, C, D,* and *E*.

We illustrate the analysis of variance of a 2^{5-2} fractional factorial design with a numerical example.

Example 18.6.3 (Analyzing a one-quarter replication of a 2^5 design using Yates's algorithm) *Using generators ABE and CDE, a* 2^{5-2} *fractional factorial experimental design yielded the data shown in Table 18.6.8. Assuming that the necessary interaction effects are negligible, we want to estimate all the main effects and use a normal probability plot to verify if any of the effects is significant. All eight runs were carried out in random order.*

Solution: We can ignore the letters D and E and apply Yates's algorithm for the 2^3 factorial design to obtain Table 18.6.9. Now the normal probability (see Section 5.8) plot

Table 18.6.7 The 2^{5-2} fractional factorial design with design generators *ABE* and *CDE*.

A	B	C	$D = ABC$	$E = AB$	Using other notation
$-$	$-$	$-$	$-$	$+$	*e*
$+$	$-$	$-$	$+$	$-$	*ad*
$-$	$+$	$-$	$+$	$-$	*bd*
$+$	$+$	$-$	$-$	$+$	*abe*
$-$	$-$	$+$	$+$	$+$	*cde*
$+$	$-$	$+$	$-$	$-$	*ac*
$-$	$+$	$+$	$-$	$-$	*bc*
$+$	$+$	$+$	$+$	$+$	*abcde*

Table 18.6.8 Results of an experiment using a 2^{5-2} fractional factorial design to improve the performance of a packaging machine.

e	*ad*	*bd*	*abe*	*cde*	*ac*	*bc*	*abcde*
3.84	3.20	3.15	3.75	3.95	4.12	3.72	4.69

Table 18.6.9 ANOVA table of the 2^{5-2} fractional factorial design.

Treat-ments	Obser-vations	(1)	(2)	(3)	Effects [(3)/4]	SS [$(3)^2/8$]	Aliases
(e)	3.84	7.04	13.94	30.42	3.8025*		$I + ABE + CDE + ABCD$
$a(d)$	3.20	6.90	16.48	1.10	0.275	0.15125	$A + BE + ACDE + BCD$
$b(d)$	3.15	8.07	-0.04	0.20	0.05	0.00500	$B + AE + BCDE + ACD$
$ab(e)$	3.75	8.41	1.14	2.04	0.51	0.52020	$AB + E + ABCDE + CD$
$c(d)(e)$	3.95	-0.64	-0.14	2.54	0.635	0.80645	$C + ABCE + DE + ABD$
ac	4.12	0.60	0.34	1.18	0.295	0.17405	$AC + BCE + ADE + BD$
bc	3.72	0.17	1.24	0.48	0.12	0.02880	$BC + ACE + ADE + AD$
$abc(d)(e)$	4.69	0.97	0.80	-0.44	-0.11	0.02420	$ABC + CE + ABDE + D$

*This entry is equal to (3)/8.

in Figure 18.6.2 shows that the estimates of all effects (Table 18.6.9) fall on a straight-line. This implies that *none* of the effects are significantly different from zero.

PRACTICE PROBLEMS FOR SECTION 18.6

1. Refer to the experiment in Problem 1 of Section 18.4. Suppose that due to some lab constraints, only eight runs could be done. Construct an appropriate design and carry out the statistical analysis, using the data from Problem 1 of Section 18.4.

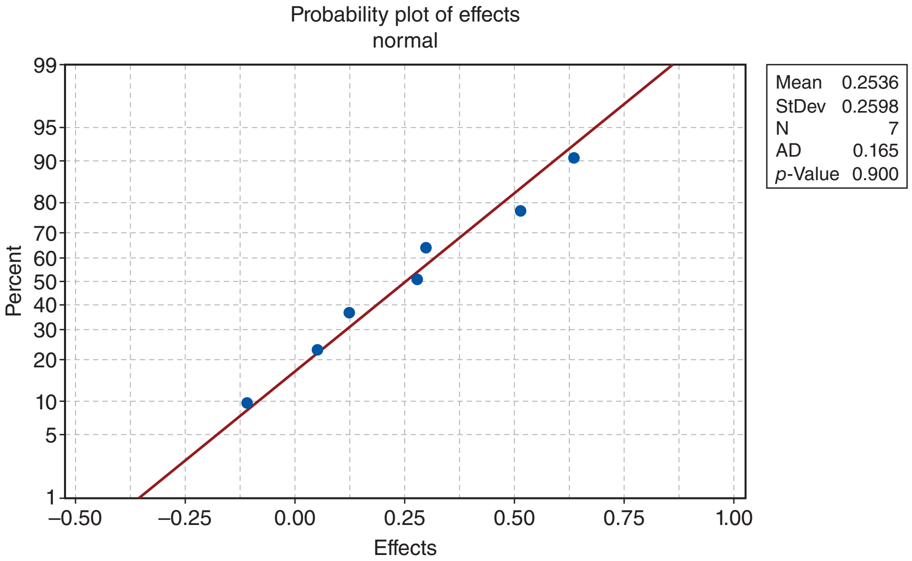


Figure 18.6.2 MINITAB printout of the normal probability plot of estimates of effects (Table 18.6.9).

2. Refer to Problem 2 of Section 18.4. Suppose that we are interested in running only a one-half fraction of the 2^4 design, that is, a design with 2^{4-1} treatments (in the four factors A, B, C, D) with two replicates, so that $N = 2 \times 2^{4-1} = 16$. Use the data in Problem 2 of Section 18.4 to form an appropriate Y column of the 2^{4-1} design, and analyze this one-half fraction replicated twice. Construct the normal probability plot of the effects and determine which effects are significant.
3. Suppose that in Problem 4 of Section 18.4 we find that factor D is unimportant. Construct a 2^3 full factorial design using the remaining three factors and then reintroduce the factor D. This gives a 1/2 replication of the 2^4 design. Analyze the data as a 1/2 replication of the 2^4 design using the appropriate observations from Problem 4 of Section 18.4 as the new data.
4. Construct a 2^{5-1} design using the five-factor interaction as the generator ($I = ABCDE$). Consider the *appropriate* observations from Problem 4 in Section 18.5 as the data for the 2^{5-1} design. Analyze the data showing that effects are confounded with each other.
5. Construct the normal probability plot of the effects in Problem 4 and determine which effects are significant. Now pool the sums of squares corresponding to the nonsignificant effects and use the pooled sum of squares to estimate the error variance σ^2. Use this estimate of the error variance to conduct the testing of the usual hypotheses for the remainder of the effects at the 5% significance level. Compare the conclusions you obtained using normal probability paper and the ANOVA table.
6. Construct a 2^{5-2} fractional factorial design by selecting two independent generators so that no two-factor interaction is their generalized interaction. Consider the appropriate observations from Problem 4 of Section 18.5 as the data for the 2^{5-2} design. Analyze the data, giving the complete alias structure for this design. Construct the normal probability plot of the effects and determine which effects are significant.

7. Refer to Problem 6. Assume that the two-factor and higher-order interactions are equal to zero, and pool the corresponding sums of squares to estimate the error variance σ^2. Use this estimate of the error variance to conduct the usual testing of hypotheses for the main effects at the 5% significance level, and compare the results to the conclusions made in Problem 6, using a normal probability plot of the effects.
8. In a study to determine the compressive strength of cylinders of concrete, five variables—type of sand, type of cement, amount of water, time to mix, and time in mold—were simultaneously studied using a 2^{5-1} fractional factorial design, with generator $-ABCDE$. The (coded) responses displayed below are in the random sequence in which the experiments were performed. (a) Construct the normal probability plot of the effects to determine the effects having the greatest influence on the compressive strength. (b) Pool the sums of squares corresponding to the non-significant effects to determine an estimate of the variance σ^2. (c) Determine 95% intervals for the effects that are found to be significant.

A	B	C	D	E	y
−	+	+	−	−	10.2
+	−	−	+	−	17.5
−	−	+	+	−	13.0
−	+	+	+	+	17.3
−	−	−	−	−	13.4
−	−	+	−	+	20.2
+	−	+	−	−	18.1
+	+	−	+	+	15.7
+	+	+	+	−	15.1
−	+	−	+	−	10.6
+	−	+	+	+	19.1
−	+	−	−	+	16.9
+	+	−	−	−	14.8
−	−	−	+	+	19.5
+	+	+	−	+	15.7
+	−	−	−	+	19.2

9. Refer to Problem 8. Construct a 2^3 full factorial design using factors A, B, and E, that is, ignoring factors C and D, so that the 2^{5-1} fractional factorial design becomes a 2^3 replicated design. Then, analyze the data in Problem 8, regarding the data as arising from a 2^3 replicated design, and compare the results to those in Problem 8.

18.7 CASE STUDIES

Case Study 1 (Data on a potato crop from an experiment carried out at Wimblington, UK)[2].

[2] **Source:** Yates, 1958. Used with permission.

Table 18.7.1 Potato crop data.

Replication I		Replication II		Replication III		Replication IV	
Block I	Block II	Block I	Block II	Block I	Block II	Block I	Block II
1(101)	n(106)	1(106)	n(89)	1(87)	n(128)	1(131)	n(103)
nk(291)	k(265)	nk(306)	k(272)	nk(334)	k(279)	nk(272)	k(302)
nd(373)	d(312)	nd(338)	d(324)	nd(324)	d(323)	nd(361)	d(324)
kd(398)	nkd(450)	kd(407)	nkd(449)	kd(423)	nkd(471)	kd(445)	nkd(437)
1163	1133	1157	1134	1168	1201	1209	1166

This experiment was carried out to study the effects of three fertilizers—nitrogen (n), potash (k), and dung (d)—on potato crops. The response variable is the yield (lbs) of potatoes per plot (1/60 acre). The experiment is replicated four times. The eight treatment combinations consisted of using

$$\begin{array}{ccccc} \text{Sulphate of ammonia } (n) & & \text{Sulphate of potash } (k) & & \text{Dung } (d) \\ \left\{\begin{array}{c} \textit{None} \\ 0.45 \text{ cwt N per acre}\end{array}\right\} & \times & \left\{\begin{array}{c} \textit{None} \\ 1.12 \text{ cwt K}_2\text{O per acre}\end{array}\right\} & \times & \left\{\begin{array}{c} \textit{None} \\ 8 \text{ tons per acre}\end{array}\right\} \end{array}$$

where 1 cwt = 100 lbs. The original experiment was carried out in blocks of eight plots each, but we present the data in blocks consisting of four plots each. The data are given in Table 18.7.1 and the last row shows the block totals.

(a) Write down, in words, an explanation of the proposed design for the agronomist.
(b) Write down the mathematical model you would need to analyze these data.
(c) Do the complete analysis, including the testing of all the usual hypotheses.
(d) Interpret the results you obtained in part (c).

Case Study 2 (Eddy current probe sensitivity study[3]) The data prepared for this case study are a subset from a study conducted by Capobianco, Splett, and Iyer.

The goal of the project was to develop a nondestructive portable device for detecting cracks and fractures in metals. A primary application would be the detection of defects in airplane wings. The internal mechanism of the detector would be used for sensing crack-induced changes in the detector's electromagnetic field, which would in turn result in changes in the impedance level of the detector. This change of impedance is termed sensitivity, and it is a subgoal of this experiment to maximize such sensitivity as the detector is moved from an unflawed region to a flawed region on the metal.

There were three detector wiring component factors under consideration:

1. X_1 = number of wire turns
2. X_2 = wire winding distance
3. X_3 = wire gauge

[3] Source: *NIST and SEMATECH (2003)*

Since the maximum number of runs that could be afforded in time and cost was $n = 10$, a 2^3 full factorial experiment (involving $n = 8$ runs) was chosen. With an eye to the usual monotonicity assumption for two-level factorial designs, the selected settings for the three factors were as follows:

1. $X_1 =$ number of wire turns: 90 coded to -1; 180 coded to $+1$.
2. $X_2 =$ wire winding distance: 0.38 coded to -1; 1.14 coded to $+1$.
3. $X_3 =$ wire guage: 40 coded to -1; 48 coded to $+1$.

The experiment was executed in completely random order, and the data obtained are given in Table 18.7.2.

(a) Write down the mathematical model you would need to analyze this set of data.
(b) Do the complete analysis, including the testing of all the hypotheses of interest and a residual analysis. Interpret the results.
(c) Estimate the effects of interest.

Table 18.7.2 Eddy current probe sensitivity data.

Y Probe impedance	X_1 Number of turns	X_2 Winding distance	X_3 Wire gauge	Run
1.70	−1	−1	−1	2
4.57	+1	−1	−1	8
0.55	−1	+1	−1	3
3.39	+1	+1	−1	6
1.51	−1	−1	+1	7
4.59	+1	−1	+1	1
0.67	−1	+1	+1	4
4.29	+1	+1	+1	5

18.8 USING JMP

This section is not included in the book, but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. Penicillin production requires a fermentation step that must be done in batches. One difficulty in the producing of successive batches is that the nutrient and corn steep liquor vary. A study was begun to determine whether changes in temperature and pH might increase the penicillin yields for a new set of fermenters. A 2^2 factorial design was employed and a new batch of corn steep liquor was used for each set of four runs. Use $\alpha = 0.05$. The results shown below were obtained:

Design		Penicillin yields				
(pH)	(Temperature)	(Corn steep liquor batches)				
A_1	A_2	1	2	3	4	5
−1	−1	40	35	28	27	33
1	−1	95	80	94	76	83
−1	1	66	50	48	45	61
1	1	124	98	105	96	100

(a) Write out a suitable analysis of variance model.
(b) Test the hypothesis that there are no treatment differences.
(c) Test the hypothesis that there are no differences between the corn steep liquor batches.
(d) Determine the pH and temperature effects.
(e) Separate individual degree of freedom sums of squares for these effects in the analysis of variance table.
(f) Estimate σ^2 and make a 95% interval statement for the pH × temperature interaction effect.

2. In the study of the flooding capacity of a pulse column, a 2^3 factorial design was employed in which two pulse amplitudes, two frequencies of pulsation, and two levels of flow ratio were varied. The whole experiment was replicated twice. The results obtained are given below.
(a) Estimate the factorial effects.
(b) Estimate the variance of estimates of factorial effects.
(c) Make a 95% interval statement for each of the various effects.
(d) Could any of the variables—amplitude, frequency, or flow ratio—be considered to be unimportant in affecting the response over the experimental region?
(e) Consider the case that second column of observations were randomly run by a different operator. Reestimate the variance and determine whether there is a statistically significant operator-to-operator difference.

Amplitude	Frequency	Flow ratio	Capacity	
−1	−1	−1	179	184
1	−1	−1	330	338
−1	1	−1	280	297
1	1	−1	300	312
−1	−1	1	185	187
1	−1	1	288	304
−1	1	1	251	271
1	1	1	193	198

3. To determine the best production characteristics for a new method of manufacturing adiponitile (ADN), considerable development work was required on the purification system. The following replicated 2^3 factorial design was performed to study the effects

of A_1: ADN feed rate; A_2: solvent-to-feed ratio; and A_3: temperature, on the response variable ADN purity. Each replicate represented one week's work. The trials were performed randomly within each week for a total of three weeks. The result obtained are given in the table below:

			Response y in Replicates		
A_1	A_2	A_3	1	2	3
−	−	−	2.58	2.66	2.74
+	−	−	3.04	2.96	3.26
−	+	−	2.81	2.63	3.07
+	+	−	3.12	3.17	3.28
−	−	+	2.45	2.49	2.65
+	−	+	2.65	2.62	2.81
−	+	+	2.45	2.54	2.67
+	+	+	2.74	2.72	3.00

(a) Using Yates's algorithm, estimate the factorial effects.
(b) Construct an appropriate analysis of variance table.
(c) Suppose that each of the three replicates represents a block. Comment on the hypothesis that there are no statistically significant differences between the blocks.
(d) Test the hypothesis that the contrast of block 3 versus blocks 1 and 2 equals zero.
(e) Make a 95% interval statement for the effect of "ADN feed rate," that is, for the A_1 effect.

4. In the manufacture of chemical products by electrolysis, the product yields depend on A_1: current density; A_2: cathode configuration; and A_3: the flow rate of the catholyte. The following 2^3 factorial design, blocked into groups of four experiments to eliminate day-to-day effects, was run to study the effects of these variables on the process yields y. The data obtained are shown below:

Block 1				Block 2				Block 3				Block 4			
A_1	A_2	A_3	y	A_1	A_2	A_3	y	A_1	A_2	A_3	y	A_1	A_2	A_3	y
−	−	−	14.6	−	−	+	19.3	−	−	+	21.3	−	−	−	21.7
+	−	+	17.4	+	−	−	16.4	+	−	−	19.5	+	−	+	22.5
−	+	+	17.4	−	+	−	15.2	−	+	−	17.6	−	+	+	24.0
+	+	−	10.2	+	+	+	16.0	+	+	+	18.3	+	+	−	16.5

(a) Using Yates's algorithm, estimate the effects of the controlled variables.
(b) Using an analysis of variance table, estimate σ^2.
(c) Is there a statistically significant linear time trend between the four blocks? Specify α.

5. Micro miniature integrated circuits are manufactured, in part, by depositing thin films of dielectric material in predesigned patterns. Prior to mounting the film, the substrate must be prepared. In a study to determine the best operating condition for its preparation, four variables were studied. The experiment was blocked into two

blocks of eight runs each, each block corresponding to a day. The design and the response (thickness of deposited SiO film in 1000 A) are shown below. Note that each replication represents a principal block generated by the same defining contrast.

(a) Estimate the effects and determine, using normal probability paper, the statistically significant effects.

(b) Construct the analysis of variance table and estimate σ^2.

(c) Test the hypothesis that there is no difference between the block means. Specify α.

Variables	(−)	(+)
A_1, substrate temperature,°C	250	300
A_2, vacuum in chamber 1, mm of Hg	1×10^{-5}	1.5×10^{-5}
A_3, vacuum in chamber 2, mm of Hg	1×10^{-5}	1.5×10^{-5}
A_4, pattern type	Type A	Type B

Block I					Block II				
A	B	C	D	y	A	B	C	D	y
−	−	−	−	3.43	−	−	−	−	3.62
+	−	−	+	4.04	+	−	−	+	4.17
−	+	−	+	3.57	−	+	−	+	3.48
+	+	−	−	3.86	+	+	−	−	4.18
−	−	+	+	4.09	−	−	+	+	3.27
+	−	+	−	3.27	+	−	+	−	4.35
−	+	+	−	3.15	−	+	+	−	4.09
+	+	+	+	4.20	+	+	+	+	3.52

6. Refer to Problem 5. Reanalyze only the data provided by Block I of the previous problem. What is the generator of this fractional factorial design?

7. Construct a 2^{6-1} fractional factorial design using the six-factor interaction as the design generator. Write down the complete alias structure.

8. Construct a 2^{6-2} fractional factorial design by using two four-factor interactions as the design generators such that their generalized interaction is another four-factor interaction. What is the complete defining relation? Write down the complete alias structure.

9. In Problem 5, instead of two blocks, treat the data as two replications of a 2^{4-1} fractional factorial design. Estimate the permissible main effects and interaction effects. Assuming all three-factor and higher-order interactions to be zero, conduct a test of hypothesis for each main effects to be zero. Is it possible to conduct the testing of a hypothesis for some of the two-factor interaction effects? If not possible, explain why not. Use $\alpha = 0.05$.

10. In Problem 9, construct the normal probability plot, and determine which effects appear to be significant. Compare the results obtained in Problem 9. Perform residual analysis to verify that the model is adequate.

11. An engineer from the chemical industry studied effects of four factors on the yield of a chemical. The four factors are A (amount of catalyst), B (temperature), C (pressure), and D (reaction time). Each factor is at two levels. The run order and the yields are as shown below:

Run order	A	B	C	D	Yield (lb)	Run order	A	B	C	D	Yield (lb)
7	−	−	−	−	15	4	−	−	−	+	20
9	+	−	−	−	18	10	+	−	−	+	27
16	−	+	−	−	23	11	−	+	−	+	21
12	+	+	−	−	17	2	+	+	−	+	17
3	−	−	+	−	16	5	−	−	+	+	22
8	+	−	+	−	25	15	+	−	+	+	29
1	−	+	+	−	19	13	−	+	+	+	18
6	+	+	+	−	24	14	+	+	+	+	26

(a) Estimate the effects and determine, using normal probability paper, the statistically significant effects.

(b) Construct the ANOVA table for these data. In this problem is it possible to estimate the error variance σ^2?

12. Two breeds of rabbits (R_1, R_2) are fed two types of diet (D_1, D_2) supplemented with two levels of proteins (P_1, P_2) using two methods (M_1, M_2), liquid and solid. The weight gain in grams by each rabbit at the end of each period was recorded and the data obtained is shown below. The experiment was carried out in a random order, and each experiment was replicated twice.

	R_1				R_2			
	D_1		D_2		D_1		D_2	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
M_1	367	483	407	411	449	423	393	409
	417	397	402	393	429	419	379	401
M_2	349	369	437	371	305	353	381	379
	473	359	362	413	389	429	399	411

(a) Analyze these data and determine which effects are significant at the 5% level of significance.

(b) Perform the residual analysis and examine if the model used is adequate.

13. Refer to Problem 12. Analyze the data of Problem 12 as a 2^4 experiment by using only the first replication, which is recorded on the top line of each cell. Construct the normal probability plot of the effects, and check which main effects or interaction effects seem to be significant.

14. In Problem 13, estimate the error variance using the sum of squares corresponding to the effects that seemed to be insignificant. Then conduct testing of hypotheses for the remainder of the effects. Use $\alpha = 0.05$.

15. An agronomist planned an experiment to study the effects of some fertilizers on the pea crop. The fertilizers included four factors, each at two levels: M (manure), N (nitrogen), P (phosphorus), and K (potassium). The experiment was conducted in random order and replicated twice. The yield (lb) of peas per plot in each replication is as shown below.

Treat.	1	*m*	*n*	*mn*	*p*	*mp*	*np*	*mnp*	*k*	*mk*	*nk*	*mnk*	*pk*	*mpk*	*npk*	*mnpk*
Rep. I	75	69	55	59	73	85	49	63	55	67	72	84	76	69	65	63
Rep. II	68	63	64	73	66	78	57	69	68	72	69	78	70	66	59	57

Analyze these data and determine which effects are significant at the 5% level of significance. Estimate the error variance σ^2.

16. In Problem 15, using variables that are found significant as independent variables, write down the regression model for yield of peas. Can you use this regression model to predict the future yield?

17. In Problem 15, while planning the experiment, the agronomist had used two depths to sow the seeds and two widths between the plants. At the end of the experiment, she felt that the depth of seeds and the width between the plants had some significant effect on the yield, but due to lack of resources and time, the agronomist decided not to conduct another experiment.
    (a) Using the data in Problem 15, consider the data as resulting from a 2^{6-2} fractional factorial experiment that employed two four-factor interactions as design generators, and reanalyze the data.
    (b) Write down the complete defining relation, and then use it to give the complete alias structure.

18. A 2^{4-1} experiment was conducted using the four-factor interaction $ABCD$ as the design generator. The data obtained from this experiment were as shown below:

1	*ab*	*ac*	*ad*	*bc*	*bd*	*cd*	*abcd*
19	25	29	20	28	31	33	37

    (a) Estimate the various effects and then prepare a normal probability plot of these effects.
    (b) Find the error mean sum of squares by collapsing the sum of squares corresponding to the effects that you find in part (a) are not insignificant. Use this error mean sum of squares to test the usual null hypothesis for each main effect. Use $\alpha = 0.05$.

19. Assume that in Problem 12, the experimenter was only able to run a 2^{4-1} design, with one replication, because only four rabbits of each breed were available. Give a plan for this experiment which is such that we can estimate all the main effects clearly, under

the assumptions that the three-factor interactions are zero. What other assumption is needed to test whether these effects are significant at the 5% level of significance?

20. Give a design that would split a 2^6 design into 2^3 blocks each of 2^3 units using the defining contrasts $ABCD$, $CDEF$, $ACDE$.

21. In Problem 20, use the defining contrasts ABC, DEF, $ABCD$. Comment on the choice of this set of defining contrasts.

22. In an experiment for studying the production of a chemical, four factors, A, B, C and D were used, each at two levels. The experiment was completely randomized, and the following data was obtained.

    $(1) = 20.5,\ a = 24.6,\ b = 19.8,\ c = 21.4,\ d = 23.2,\ ab = 21.7,\ ac = 20.5,\ ad = 23.7,$
    $bc = 21.9, bd = 25.4,\ cd = 22.5,\ abc = 22.5,\ abd = 23.4,\ acd = 25.4,\ bcd = 21.5,$
    $abcd = 24.7.$

    Select eight observations carefully from these data in such a way, that if assuming three-factor interactions are zero, one can estimate all the main effects. Is it possible to estimate all the main effects by selecting only four observations and assuming two-factor and three-factor interactions are zero?

23. The product yields of an unwanted by-product were measured (in percentages) for two different catalysts C_1, C_2 each at two different pressures P_1, P_2. The experiment was carried out by two different analysts A_1, A_2 at two different labs L_1, L_2. The whole experiment was replicated twice. The results obtained are shown below. Estimate the main effects and the interactions. Determine which factors have significant effects (if any) on the yield of the unwanted by-product. Since the experiment is replicated twice, check if any of the interactions are also significant at the 5% level of significance.

	C_1				C_2			
	P_1		P_2		P_1		P_2	
	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
L_1	37	27	49	29	57	61	53	43
	43	47	53	26	62	51	39	43
L_2	57	42	63	59	38	51	61	54
	63	49	54	41	46	59	64	57

24. In a finishing process for metal plates, the smoothness of the surface is very important. The factors responsible for the smoothness are solution temperature T, solution concentration C, size of the plate S, and, finally, the tension of the metal M. In order to study the effects of these factors on smoothness, an experiment was conducted using the four factors, each at two levels. The observations presented below are the scores, which represent the percentage of the desired smoothness. The experiment was completely randomized.

	T_1				T_2			
	C_1		C_2		C_1		C_2	
	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
M_1	79	81	83	76	75	83	87	82
M_2	57	42	63	59	38	51	61	54

(a) Estimate all the main effects and interactions.

(b) Making the necessary assumptions, test at the 1% level of significance if any of the main effects or two-factor interactions is significantly different from zero.

Chapter 19

RESPONSE SURFACES

The focus of this chapter is the development of first-order and second-order (central composite) designs employed for fitting response surfaces.

Topics Covered

- First-order designs
- Second-order designs
- Central composite designs (CCDs)
- Some other first-order and second-order designs
- Determination of the optimum or near optimum point
- The method of steepest ascent
- Analysis of a fitted second-order response surface

Learning Outcomes

After studying this chapter, the reader will be able to

- Select appropriate designs to fit first-order and second-order models.
- Use the least-squares method to fit a desired model.
- Use analysis of variance techniques to verify the adequacy of the fitted models.
- Analyze the fitted response surface to determine the nature of the response surface.
- Use certain techniques to determine the optimum (or near-optimum) point of the response surface.
- Use statistical packages MINITAB, R, and JMP to analyze response surface.

19.1 INTRODUCTION

Response surface methodology (RSM) allows an experimenter to explore an unknown functional relationship between a response variable Y and k controlled or independent

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

variables, say $\xi_1,\ \xi_2,\ \ldots,\ \xi_k$. The method was introduced by Box and Wilson (1951). The main goals of RSM are to use a sequence of planned experiments to seek an optimal response and to assess a functional relationship in the neighborhood of the optimal response. In this chapter, we study various commonly used response surface designs and their analysis.

19.1.1 Basic Concepts of Response Surface Methodology

Suppose an unknown functional relationship relating the response Y (an overall yield or percentage of an unwanted by-product of a chemical process, for example) with the levels of k controlled variables $\xi_1, \xi_2, \ldots, \xi_k$ (temperature, pressure, amount of catalyst, reaction time, for example) is

$$Y = f(\xi_1, \xi_2, \ldots, \xi_k) + \varepsilon \tag{19.1.1}$$

where ε represents the random experimental error due to some unknown or uncontrollable variables. As usual, it is assumed that ε is normally distributed with mean zero and variance σ^2. Thus, taking the expected value of Y in (19.1.1), we obtain

$$E(Y) = \eta = f(\xi_1, \xi_2, \ldots, \xi_k) \tag{19.1.2}$$

over an experimental region defined by acceptable ranges of the k controlled variables $\xi_1, \xi_2, \ldots, \xi_k$.

In this chapter, we have two goals: (i) find a polynomial model that can adequately estimate the functional relationship given by (19.1.2) in the experimental region of the controlled variables, and (ii) determine an optimal point, that is, the point in the experimental region (defined by the ξ_i's) at which $f(\xi_1, \xi_2, \ldots, \xi_k)$ is optimal. For example, interest may lie in finding where the yield is maximized or the percentage of waste or unwanted byproduct is minimized.

If the function is continuous over a region, which is relatively small, then it may be usefully approximated by a first-order Taylor series about $\xi_0 = (\xi_{10}, \xi_{20}, \ldots, \xi_{k0})$, a selected point within the region. The Taylor series takes the form

$$\eta = f(\xi_{10}, \xi_{20}, \ldots, \xi_{k0}) + (\xi_1 - \xi_{10})\frac{\partial f}{\partial \xi_1} + (\xi_2 - \xi_{20})\frac{\partial f}{\partial \xi_2} + \cdots + (\xi_k - \xi_{k0})\frac{\partial f}{\partial \xi_k} \tag{19.1.3}$$

where each of the derivatives is evaluated at the point ξ_0, where $\xi_0 = (\xi_{10}, \xi_{20}, \ldots, \xi_{k0})$. Then to good approximation, the model (19.1.3) may be rewritten as

$$\eta = \gamma_0 + \gamma_1\xi_1 + \gamma_2\xi_2 + \cdots + \gamma_k\xi_k = \gamma_0 + \sum_{i=1}^{k} \gamma_i\xi_i \tag{19.1.4}$$

where

$$\gamma_0 = f(\xi_{10}, \xi_{20}, \ldots, \xi_{k0}) - \xi_{10}\frac{\partial f}{\partial \xi_1} - \xi_{20}\frac{\partial f}{\partial \xi_2} - \cdots - \xi_{k0}\frac{\partial f}{\partial \xi_k} \tag{19.1.5}$$

with

$$\gamma_i = \frac{\partial f}{\partial \xi_i}, \quad i = 1, 2, \ldots, k \tag{19.1.6}$$

Each of the derivatives (19.1.6) is evaluated at the point ξ_0. The coefficients γ_i are usually called the first-order coefficients. The model in (19.1.3) may be viewed as a first-order approximating polynomial. If this first-order model should prove inappropriate to represent the response function, a second-order Taylor series is often employed. The second-order polynomial approximation is

$$\eta = \gamma_0 + \sum_{i=1}^{k} \gamma_i \xi_i + \sum_{i=1}^{k} \gamma_{ii} \xi_i^2 + \sum_{i,j=1} \sum_{i<j} \gamma_{ij} \xi_i \xi_j \tag{19.1.7}$$

where γ_{ii} is the quadratic coefficient of the quadratic term ξ_i^2, and γ_{ij} is the cross product or two-factor interaction coefficient between variables ξ_i and ξ_j. Together, the k quadratic and $k(k-1)/2$ cross-product coefficients determine the second-order portion of the polynomial model. Note that $\gamma_{ij} = \dfrac{\partial^2 f}{\partial \xi_i \partial \xi_j}$, where these derivatives are evaluated at the point ξ_0.

In discussing both the experimental designs and the analysis of the data, one can code u, where u indexes the settings of the variable ξ_i, say ξ_{iu}, $u = 1, \ldots, n$, by standardized variables x_{iu} defined by

$$x_{iu} = \frac{\xi_{iu} - \xi_{i0}}{c_i} \tag{19.1.8}$$

Here ξ_{i0} is the midpoint of the experimental region with respect to ξ_i and its settings ξ_{iu}, $u = 1, \ldots, n$, and where c_i is some convenient scale factor chosen so that x_{iu} are convenient numbers, easy to work with. For instance the variables ξ_i with two settings are easily coded to $x_{i1} = 1$, $x_{i2} = -1$. As an example, if the two settings (i.e. u =1, 2) of a variable ξ_{iu} are 12 and 20, then the standardized variable x_{iu} could be defined as

$$x_{iu} = \frac{\xi_{iu} - 16}{4}$$

Here $x_{i1} = -1$, and $x_{i2} = 1$, since c_i is chosen to be 4.

The first-order model may now be written as

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \tag{19.1.9}$$

and the second-order model

$$\eta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i,j=1} \sum_{i<j} \beta_{ij} x_i x_j \tag{19.1.10}$$

where β_0 is a constant, β_i are the k first-order coefficients, β_{ii} are the k quadratic coefficients, and β_{ij} are the $k(k-1)/2$ cross-product or interaction coefficients for the models written in terms of the x_i. Figure 19.1.1a,b show in three dimensions the response surface plots of the first-order and second-order response surfaces in $(X_1, X_2, \hat{Y})$ space, and Figure 19.1.2a,b show the contour plots of the first-order and second-order response surfaces (see Example 16.2.1).

If the experimental region is small, then a first-order model is often useful for arriving at the point of an optimal response. Also, a second-order model usually is quite adequate in illustrating the functional relationship between the response and the controlled variables. For example, from the contour plots (where contours are defined by setting

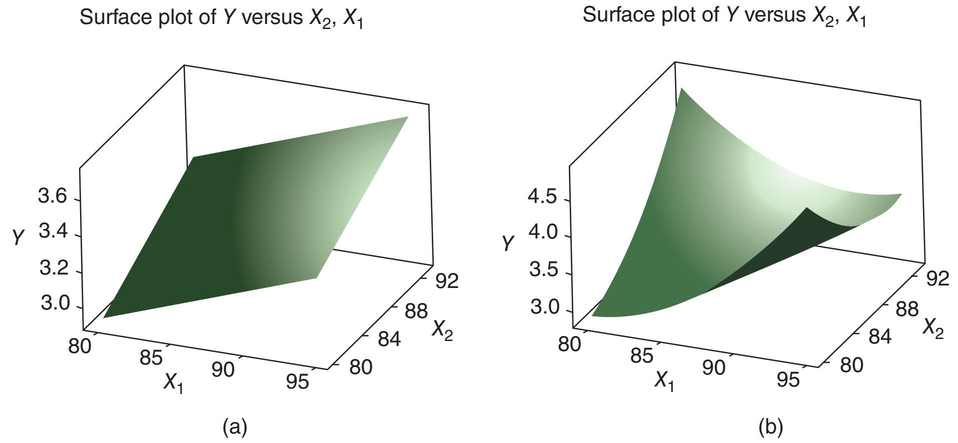


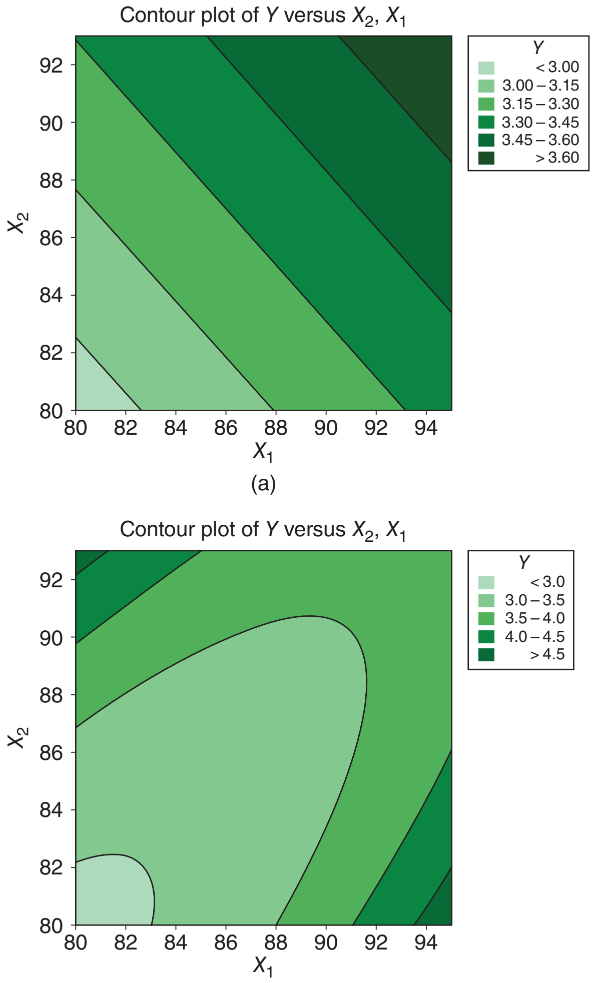
Figure 19.1.1 (a) The surface plot for the fitted model $\hat{Y} = -1.66 + 0.0283X_1 + 0.0290X_2$. (b) The surface plot for the fitted regression model $\hat{Y} = -7.679 + 0.635X_1 - 0.447X_2 + 0.005X_1^2 + 0.011X_2^2 - 0.016X_1X_2$.

$\hat{y} = 3.00, 3.15, 3.30, 3.45$, and 3.60) of the first-order model given in Figure 19.1.2a, which consists of parallel lines, we find the direction, usually called the *steepest ascent* or *steepest descent* (to be discussed later), that allows us to maximize the yield or minimize the percentage of waste or unwanted byproduct. That is, we will be led to the point where the response is optimized. The contour plots (where contours are defined by setting $\hat{y} = 3.00, 3.50, 4.00$, and 4.50) of the second-order model given in Figure 19.1.2b, called a *rising ridge*, indicates the direction where the response can be optimized.

Figure 19.1.3 shows some other contour plots of the second-order model that are commonly encountered in RSM.

The contours in Figure 19.1.3 indicate the following: (a) a point of maximum response exists in the experimental region; (b) the response increases or decreases as we move away from the center of the experimental region, where the increase or decrease in response depends upon the direction in which we move; (c) there is more than one optimal point, and these occur anywhere on a line located in the experimental region; (d) the optimal point is far removed from the experimental region (in this case, to find the [optimal] point that maximizes the response, further exploration is necessary); (e) the optimal point is, again, far removed from the experimental region (in this case, to find the [optimal] point that minimizes the response, further exploration is necessary); and (f) a point of minimum response exists in the experimental region.

Suppose now that the experimenter at the early exploratory stage of an experiment feels that in order to adequately describe the functional relationship given by (19.1.2), he/she needs to fit a second-order polynomial instead of a first-order polynomial in the variables $x_1, x_2, \ldots, x_k$. Then from (19.1.9) to (19.1.10), we note that the number of parameters $\beta_0, \beta_1, \beta_2, \ldots$ that he/she would need to estimate increases rapidly as we move from a first-order to a second-order polynomial. Since fitting a polynomial requires the number of observations to be at least as large as the number of unknown parameters, the number of necessary experiments to be carried out also increases very rapidly, particularly if the experimenter obtains only one observation from each experiment. Thus, the experimenter should be very careful in deciding on the order of the polynomial that he/she would like



(a)

(b)

Figure 19.1.2 (a) The contour plot for the fitted model $\hat{Y} = -1.66 + 0.0283X_1 + 0.0290X_2$, by setting $\hat{y} = 3.00$, 3.15, 3.30, 3.45, and 3.60. (b) The contour plot (rising ridge) for the fitted regression model $\hat{Y} = -7.679 + 0.635X_1 - 0.447X_2 + 0.005X_1^2 + 0.011X_2^2 - 0.016X_1X_2$, by setting $\hat{y} = 3.00, 3.50, 4.00, 4.50$.

to fit. Note that if the experimental region is narrow and f has a quite small curvature, then often a first-degree polynomial is considered, since the possibility of fitting such polynomials adequately is quite high. The use of first-degree polynomials is usually strongly recommended when we are at the exploratory stage of a completely new experimental situation. It often yields important information that helps in deciding future action, using a relatively small number of experiments. Indeed, it may happen that the real situation is well explained by a first-degree polynomial, so starting with a second-degree polynomial will require more resources unnecessarily.

In this chapter, we consider the problem of fitting polynomials of first-order as well as second-order. Since the fitting of a polynomial can be considered as a special case of

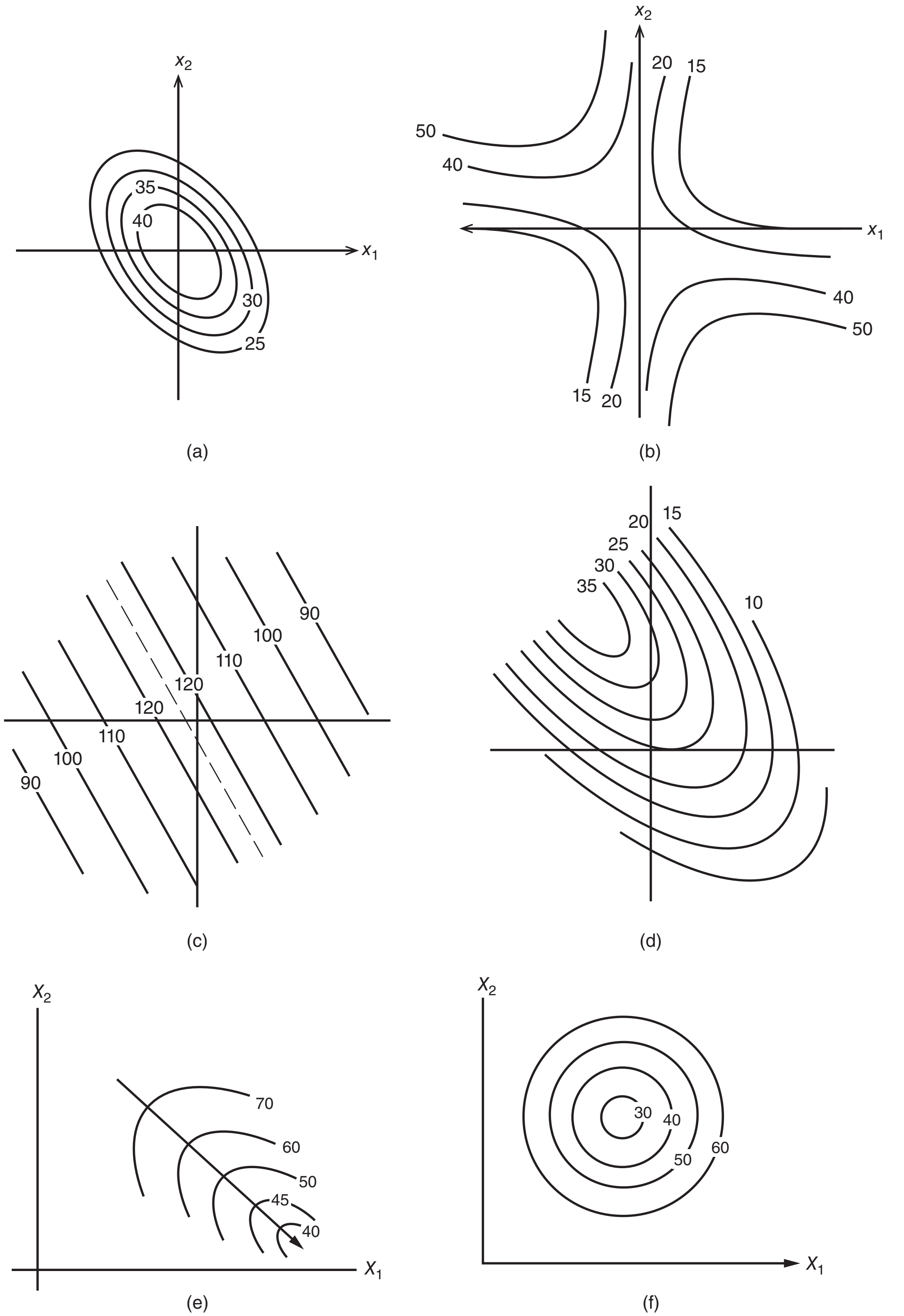


Figure 19.1.3 Illustrative contour plots provided by fitted second-order models: (a) mound, (b) saddle point, (c) stationary ridge, (d) rising ridge, (e) falling ridge, and (f) basin.

multiple linear regression, we strongly recommend that the reader review Chapter 17 on multiple linear regression. Box (1952) called designs used to fit first-order and second-order polynomials *first-order* and *second-order designs.*

19.2 FIRST-ORDER DESIGNS

Suppose now, based upon some prior information, it is felt that the response surface within the experimental region can be adequately approximated by a hyperplane, so that fit of a polynomial of the first-degree is to be made. That is, the model to be fitted is

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon \tag{19.2.1}$$

Using the data obtained by observing the response Y_i at $(X_{i1}, \ldots, X_{ik}),\ i = 1, \ldots, n$, that is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{19.2.1a}$$

where Y_i's are the responses, $\beta_0, \beta_1, \beta_2, \ldots,$ and β_k are the unknown parameters, β_0 is the intercept of this hyperplane, and $\beta_1, \beta_2, \ldots,$ and β_k are the partial regression coefficients. X_{ij} $(j = 1, 2, \ldots, k)$ is the *ith value of the controlled variable* X_j *in the ith experiment*, and ε_i's are random errors with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$. We may express (19.2.1a) in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\gamma + \varepsilon \tag{19.2.2}$$

where

$$\boldsymbol{Y}(n \times 1) = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X}[n \times (k+1)] = \begin{bmatrix} 1\ X_{11}\ X_{12} \cdots X_{1k} \\ 1\ X_{21}\ X_{22} \cdots X_{2k} \\ \vdots\ \vdots \quad \vdots \qquad \vdots \\ 1\ X_{n1}\ X_{n2} \cdots\ X_{nk} \end{bmatrix},$$

$$\boldsymbol{\gamma}[(k+1) \times 1] = \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon(n \times 1) = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Now we would like to select a design for fitting the model (19.2.2) that is best in some sense. In general, we consider the use of a design T if it allows us to estimate all the regression coefficients β_is with smallest variance. It is known that if the variables in the matrix X are functionally independent, then the requirement of smallest variance is satisfied if we chose the design T such that the matrix $X'X$ is diagonal. Thus, a class of designs that is suitable for determining the regression coefficients in the polynomial (19.2.1) is the class of orthogonal designs. A class of orthogonal designs that is of particular great interest is the class of factorial and fractional factorial designs with each factor at two levels. Moreover, if we consider a factorial design with each factor at two levels, then each x-variable in (19.2.1) takes only two values, which reduces the required number of treatments to carry out the whole experiment by a great deal. Also, as discussed before, it allows us to easily code the two levels of each factor as -1 and 1. Furthermore, this way the *center of the design*, the point representing the mid value between the high and low levels of the factors, is the origin. This property will be seen to be very useful when estimating the various regression coefficients.

Before considering some numerical examples on first-order designs, we would like to discuss some of the consequences in case the true situation is not fully described by the polynomial (19.2.1). Suppose, then, that the model (19.2.1) is not adequate, and we need to include some second-order terms. Then the new model leads to

$$Y = X\gamma + X_1\gamma_1 + \varepsilon \tag{19.2.3}$$

where, for $p_1 = k(k+1)/2$, $\gamma_1 = (\beta_{11}, \beta_{22}, \ldots, \beta_{12}, \ldots)'$ is a $[p_1 \times 1]$ vector of unknown parameters that are the coefficients of the second-degree terms $x_1^2, x_2^2, \ldots, x_k^2$, $x_1x_2, \ldots, x_{k-1}x_k$. Thus, if we fit model (19.2.1), then $\hat{\gamma}$, the least-squares the estimator of γ, is such that

$$E(\hat{\gamma}) = (X'X)^{-1}X'E(Y) \tag{19.2.4}$$

Now, if (19.2.3) is the true model, then from Equation (19.2.4) we find

$$\begin{aligned} E(\hat{\gamma}) &= (X'X)^{-1}X'(X\gamma + X_1\gamma_1) \\ &= (X'X)^{-1}(X'X)\gamma + (X'X)^{-1}X'X_1\gamma_1 \\ &= \gamma + B\gamma_1 \end{aligned} \tag{19.2.5}$$

where $\hat{\gamma} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ is the least-squares estimate of $\gamma = (\beta_0, \beta_1, \ldots, \beta_k)$ obtained when fitting (19.2.1). From this result it is obvious that $\hat{\gamma}$ is no longer an unbiased estimator of γ. In other words, each estimator $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ is a biased estimator of $\beta_0, \beta_1, \ldots, \beta_k$, respectively, where the bias is some linear combination of the terms of γ_1. The exact combination is determined by the matrix $\boldsymbol{B} = (X'X)^{-1}X'X_1$ which is also sometimes known as the *alias matrix.* The term alias, in fact, is the same as used in the discussion of confounding in Chapter 18.

Example 19.2.1 (Determining the bias) *Suppose that $k = 3$, and we consider the design T as the 1/2 replication of a 2^3 factorial experiment defined by the contrast $I = ABC$. That is,*

$$T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$$

where the upper and lower levels of each factor are coded to 1 and -1, respectively. We now show that the estimates of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ are biased if the true model consists of interaction terms as well.

Solution: In other words, we assume that the linear model is fitted, that is, we fit

$$Y = \beta_0x_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon \tag{19.2.6}$$

but the true model is

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \tag{19.2.7}$$

Thus, we have recalling that we are using the design T defined above, that

$$\gamma = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, X_1 = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix}, \text{ with } \gamma_1 = \begin{bmatrix} \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{bmatrix}$$

On using (19.2.5), the above means that

$$\begin{aligned} E(\hat{\gamma}) &= \gamma + (X'X)^{-1} X' X_1 \gamma_1 \\ &= \gamma + \left[\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \right]^{-1} \times \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1 & 1 & 1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{bmatrix} \\ &= \gamma + \frac{1}{4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 4 & 0 \\ 4 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{bmatrix} = \gamma + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{bmatrix} = \gamma + \begin{bmatrix} 0 \\ \beta_{23} \\ \beta_{13} \\ \beta_{12} \end{bmatrix} \end{aligned}$$

We now have that

$$E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1 + \beta_{23}, E(\hat{\beta}_2) = \beta_2 + \beta_{13}, E(\hat{\beta}_3) = \beta_3 + \beta_{12}$$

If we now refer back to our discussion in Chapter 18, then we note that in a one-half replication of 2^3 factorial with $ABC = I$, the alias groups are $A, BC; B, AC; C, AB$. Thus, the concept of biasedness here is the same as that of aliased groups arising in the discussion of the confounding in Chapter 18. Further, we may remark here that if we consider the further use of the design T, $I = -ABC$, a 2^{3-1} experiment, then the complete 2^3 experiment enables the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ to be be bias-free. We illustrate this with the following example.

Example 19.2.2 (Determining the bias) *Consider the use of a complete 2^3 factorial experiment when fitting (19.2.6), but where the true model is actually*

$$\begin{aligned} Y = {} & \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 \\ & + \beta_{23} x_2 x_3 + \varepsilon \end{aligned} \tag{19.2.8}$$

Here, we have

$$\gamma = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, X = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, X_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \gamma_1 = \begin{bmatrix} \beta_{11} \\ \beta_{22} \\ \beta_{33} \\ \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{bmatrix}$$

Solution: Now, on using (19.2.5), we can easily verify that

$$E(\hat{\gamma}) = \gamma + \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{22} \\ \beta_{33} \\ \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{bmatrix}$$

This implies that

$$\begin{cases} E(\hat{\beta}_0) = \beta_0 + \beta_{11} + \beta_{22} + \beta_{33} \\ E(\hat{\beta}_1) = \beta_1 \\ E(\hat{\beta}_2) = \beta_2 \\ E(\hat{\beta}_3) = \beta_3 \end{cases} \tag{19.2.9}$$

Thus, from our discussion in Examples 19.2.1 and 19.2.2, we observe that if we fit a first-order model when, in fact, the true model also contains the second-order terms, then we have the problem of biasedness. Further, it is usually true that in the exploratory stage, the experimenter does not have complete information about the real situation. It is, therefore, very important that when approximating the response surface by a first-order polynomial, we must construct the design in such a way that we can test the adequacy of the model. This is usually done by adding to a 2^k factorial or fractional factorial design, a few extra points, say n_c, at the center. By doing so, as one can see by writing down the normal equations, the values of the estimated regression coefficients $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are not changed, except that $\hat{\beta}_0$ becomes the average of all the $n + n_c$ observations. The sum of squares of deviations of the n_c responses, generated at the n_c central points from the mean of these n_c observations, provides $(n_c - 1)$ degrees of freedom for the estimation of the experimental error and one degree of freedom for estimation of the sum of coefficients of the pure quadratic terms (see (19.2.9)). The adequacy of the model is tested against the aforementioned mean square. Thus, the observations taken at the center point allow us to test the interaction terms and the pure quadratic terms. Note, however, that in the absence of the interactions, it is usually highly unlikely that the pure quadratic terms will be present in the model. For further illustration, we consider the following example.

Example 19.2.3 (A chemical production) *For a chemical production process, a chemist studied the effects of three factors, A catalyst (%), B reaction time, and C temperature, on the yield of the chemical, using a 2^3 factorial design. Table 19.2.1 shows the final results*

Table 19.2.1 Plan of a 2^3 factorial design.

x_1	x_2	x_3	Yields (y)
-1	-1	-1	20
1	-1	-1	17
-1	1	-1	18
1	1	-1	22
-1	-1	1	20
1	-1	1	19
-1	1	1	23
1	1	1	21

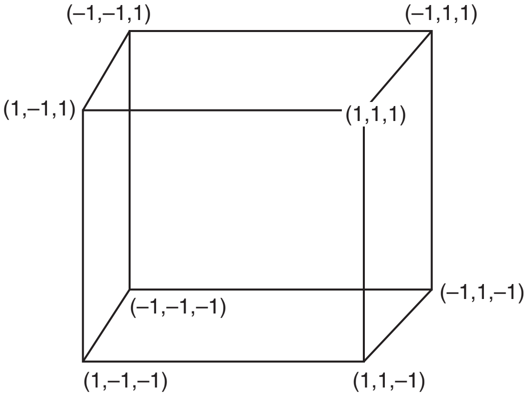


Figure 19.2.1 The 2^3 factorial runs pictured at the vertices of the unit cube in $(-1 \leq x_i \leq 1)$, $i = 1, 2, 3$.

of the experiment. All eight treatments were run in random order. The coded values of the levels are +1 and −1, which are represented by the vertices of a cube with its center at the origin (0, 0, 0), as pictured in Figure 19.2.1.

Suppose we want to fit a first-order model, that is

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{19.2.10}$$

Solution: Using the techniques of Chapter 18, we can see easily that the estimates of the regression coefficients (remember that regression coefficients are equal to one-half of the main effects and interaction) are

$$\begin{aligned}
\hat{\beta}_0 &= \frac{(y_1 + y_2 + \cdots + y_8)}{8} = 20 \\
\hat{\beta}_1 &= \frac{(-y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8)}{8} = -0.25 \\
\hat{\beta}_2 &= \frac{(-y_1 - y_2 + y_3 + y_4 - y_5 - y_6 + y_7 + y_8)}{8} = 1 \\
\hat{\beta}_3 &= \frac{(-y_1 - y_2 - y_3 - y_4 + y_5 + y_6 + y_7 + y_8)}{8} = 0.75
\end{aligned} \tag{19.2.11}$$

Table 19.2.2 ANOVA table for the data in Table 19.2.1.

Source	DF	SS	MS
$A(x_1)$	1	0.5^a	0.5
$B(x_2)$	1	8.0	8.0
$C(x_3)$	1	4.5	4.5
Residual	4	15 (by subtraction)	3.75
Total	7	28	

[a] The sum of squares due to various effects may be obtained by using the expression $SS = 2^{k-2} \times r \times (\text{estimate of effect})^2$, where r is the number of replications. For example, $SS_A = 2^{3-2} \times 1 \times (\hat{A})^2 = 2(\hat{A})^2 = 2(2\hat{\beta}_1)^2 = 2(-0.5)^2 = 0.5$, since effects are equal to two times the regression coefficients and in this example $r = 1$.

Note that $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ are the estimates of the slopes of the plane (19.2.10) in the directions of x_1, x_2, and x_3, respectively. In other words, they represent the estimated change in y per unit change in x_1, x_2, and x_3, respectively.

For testing the adequacy of the linear model, the error variance can be estimated by taking just two observations at the center of the design, i.e., (0, 0, 0). The two observations generated at the center $(0, 0, 0)$ are 20 and 21. As mentioned earlier, $\hat{\beta}_0$ is the mean of all the 10 observations, that is,

$$\hat{\beta}_0 = \frac{20 + 17 + 18 + \cdots + 21 + 20 + 22}{10} = 20.2$$

Thus, the new fitted model becomes (see (19.2.11))

$$\hat{Y} = 20.2 - 0.25x_1 + x_2 + 0.75x_3$$

An estimate of error variance (i.e. of σ^2) may be obtained from the two center points. We have that the mean of the center point is $\bar{Y}^* = \dfrac{20 + 22}{2} = 21$, so that the sum of squares of deviations for these two observations is

$$\sum (Y_i^* - \bar{Y}^*)^2 = (20 - 21)^2 + (22 - 21)^2 = 2$$

When this sum of squares is divided by its degrees of freedom, which is $2-1 = 1$, we call this result the mean square for pure error, which in this example is $2/1 = 2$ (see Table 19.2.3).

Note that we have let Y_i^* denote the ith observation at the center and $\bar{Y}^*$ denote the mean of the observations generated at the center of the design. The analysis of variance for the combined data is shown in Table 19.2.3. Note that in the following table residual degrees of freedom and the sum of squares are broken into two parts (i) lack-of-fit (ii) pure error.

If $\bar{Y}$ is the average of the observations of the noncentral points and $\bar{Y}^*$ the average of the observations at the center, the comparison $(\bar{Y} - \bar{Y}^*)$ gives an additional degree of freedom for measuring lack of fit. This is indicated by (19.2.9) and the fact that $\hat{\beta}_0 = \bar{Y}$, so that

$$E(\bar{Y}) = \beta_0 + \beta_{11} + \beta_{22} + \beta_{33}$$

Table 19.2.3 ANOVA table of the 2^3 factorial design with two points at the center.

Source	DF	SS	MS	F
$A(x_1)$	1	0.5	0.50	0.25
$B(x_2)$	1	8.0	8.00	4.00
$C(x_3)$	1	4.5	4.50	2.50
Residual	6	18.6	3.10	
Lack-of-fit	5	16.6 (by subtraction)	3.32	1.66
Pure Error	1	2.0	2	
Total	9	31.6		

Combining this with the fact that

$$E(\bar{Y}_i^*) = \beta_0 + \sum_{i=1}^{3} \beta_i(0) + \sum_{i=1}^{3} \beta_{ii}(0) + \beta_{12}(0) + \beta_{13}(0) + \beta_{23}(0) = \beta_0$$

we have that $E(\bar{Y}^*) = \beta_0$, and hence

$$E(\bar{Y} - \bar{Y}^*) = \beta_{11} + \beta_{22} + \beta_{33}$$

Now the sum of squares corresponding to this single degree of freedom is given by

$$SS_{\text{pure quadratic}} = \frac{n \times n_c}{n + n_c}(\bar{Y} - \bar{Y}^*)^2 = \frac{8 \times 2}{8 + 2}(-1)^2 = 1.6$$

where $n = 8$ and $n_c = 2$ are the number of points in the factorial part and the number of center points, respectively. Thus, the sum of squares due to lack of fit for the combined data can be obtained by subtraction, as illustrated in Table 19.2.3.

Now, from Table 19.2.3, the F-ratio for testing lack of fit is 1.66, which is insignificant at the 5% level of significance. This suggests the adequacy of the linear model. Here we can also verify whether or not there is any pure quadratic effect. We know that

$$F = \frac{MS_{\text{pure quadraric}}}{\hat{\sigma}^2} = \frac{(SS_{\text{pure quadratic}})/1}{\hat{\sigma}^2} = \frac{1.6}{2} = 0.8 < 1$$

Since the observed F-value is less than 1, there is no indication that the pure quadratic effect (curvature) should be in the model.

We may remark here that to fit a linear model, one could use a one-half or even a smaller fraction of a complete replication. Nevertheless, we point out that in the above example, we could not use even a one-half replication because in that case it would not be possible to obtain any degrees of freedom for a lack of fit term. If the number of factors is greater than 3, then it is possible to use a one-half or smaller replication of a 2^k design, depending upon the number of factors, and, by *adding n_c center points*, we can also test the adequacy of the model.

Table 19.2.4 Certain designs for $k = 2, 3$ useful for fitting first-order models.

$k = 2$		$k = 3$			$k = 3$		
x_1	x_2	x_1	x_2	x_3	x_1	x_2	x_3
−1	−1	−1	−1	1	−1	−1	−1
1	−1	1	−1	−1	1	−1	−1
−1	1	−1	1	−1	−1	1	−1
1	1	1	1	1	1	1	−1
0	0	0	0	0	−1	−1	1
·	·	·	·	·	1	−1	1
·	·	·	·	·	−1	1	1
·	·	·	·	·	1	1	1
0	0	0	0	0	0	0	0
Square (or 2^2 factorial) with center points		Tetrahedron (or simplex, or 2^{3-1} factorial) with center points			·	·	·
					·	·	·
					·	·	·
					0	0	0
					Cube (or 2^3 factorial) with center points		

Certain designs useful for fitting first-order models for $k = 2$ and 3 are given in Table 19.2.4.

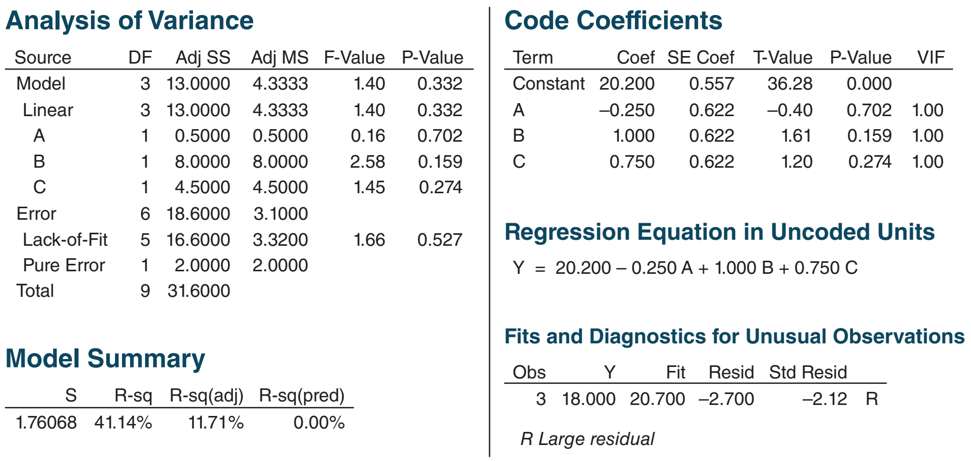
Note that all these designs contain some part of the design that could be replicated, and that an important check on the adequacy of the model is available whenever portions of the experimental design are replicated. In such cases, an estimate of σ^2 can be constructed in the usual manner at each replicated point, and when these estimates are pooled, they give an estimate of σ^2, say S^2, based on ν degrees of freedom (this was done in Example 19.2.3, where the center point was replicated, $n_c = 2$). It is possible then to partition the residual sum of squares into two portions, one portion due to intrinsic variability alone, or pure error, represented by S^2 and based on ν degrees of freedom, and a remainder portion representing the failure of the fitted responses to estimate the true response, or "lack of fit," based upon $N - k - 1 - \nu$ degrees of freedom. We explore this point further in the following example.

Example 19.2.4 (Using MINITAB and R) *Do Example 19.2.3 after adding two center points 20 and 22 using both MINITAB and R.*

MINITAB

1. Enter the factorial points (±1's) and center points (0's) in columns C1–C3.
2. Create another column, say C4, by using 1 for factorial point and 0 for center point.
3. Enter the responses in column C5.
4. From the bar menu select **Stat** > **DOE** > **Response Surface** > **Define Custom Response Surface Design.**

5. In the dialog box enter C1, C2, and C3 under Continuous Factors. Select **Low/High**, verify that in the new dialog box the **Low** column and **High** column has −1 and 1, respectively, and then click **OK**.
6. Select Designs option, and in the new dialog box, check **Specify by column** under "Point type column" and in the box next to it enter C4. Then click **OK**. Again, click **OK**. This completes the process of creating a custom design.
7. From the bar menu select **Stat** > **DOE** > **Response Surface** > **Analyze Response Surface Design.** Then enter C5 under the box Responses: appears in the new window and click **Terms** and select Linear from the pull down menu appears. Click **OK** twice. The MINITAB output appears in the **Session Window** as shown below.



```
Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
 Model           3  13.0000   4.3333     1.40    0.332
  Linear         3  13.0000   4.3333     1.40    0.332
   A             1   0.5000   0.5000     0.16    0.702
   B             1   8.0000   8.0000     2.58    0.159
   C             1   4.5000   4.5000     1.45    0.274
 Error           6  18.6000   3.1000
  Lack-of-Fit    5  16.6000   3.3200     1.66    0.527
  Pure Error     1   2.0000   2.0000
 Total           9  31.6000

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
1.76068  41.14%     11.71%       0.00%
```

```
Code Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  20.200    0.557    36.28    0.000
A         –0.250    0.622    –0.40    0.702  1.00
B          1.000    0.622     1.61    0.159  1.00
C          0.750    0.622     1.20    0.274  1.00

Regression Equation in Uncoded Units

Y = 20.200 – 0.250 A + 1.000 B + 0.750 C

Fits and Diagnostics for Unusual Observations

Obs       Y     Fit   Resid  Std Resid
  3  18.000  20.700  –2.700      –2.12  R

R Large residual
```

These results clearly match those obtained manually in Example 19.2.3.

USING R

Solution To perform the required response surface modeling in R, we can use the 'rsm()' function in the R 'library(rsm)' as shown in the following R-code. Note: here we fit a first-order model in (19.2.10) and the command 'FO(x1, x2, x3)' in rsm should be used.

```
library(rsm)
# Make a data.frame
x1 = c(-1,1,-1,1,-1,1,-1,1,0,0)
x2 = c(-1,-1,1,1,-1,-1,1,1,0,0)
x3 = c(-1,-1,-1,-1,1,1,1,1,0,0)
y = c(20,17,18,22,20,19,23,21,20,22)
data = data.frame(x1, x2, x3, y)

# Run the suggested first-order model
mod.rsm = rsm(y ~ FO(x1, x2, x3), data = data)
summary(mod.rsm)
```

```
#R summary output

              Estimate   Std. Error   t value   Pr (> |t| )
(Intercept)   20.2000    0.5568       36.28     0.0000
x1            -0.2500    0.6225       -0.40     0.7019
x2             1.0000    0.6225        1.61     0.1593
x3             0.7500    0.6225        1.20     0.2736

Multiple R-squared: 0.4114, Adjusted R-squared: 0.1171

F-statistic: 1.398 on 3 and 6 DF, p-value: 0.3318

Analysis of Variance Table

                   Df   Sum Sq   Mean Sq   F value   Pr(>F)
FO(x1, x2, x3)     3    13.0     4.3333    1.3978    0.3318
Residuals          6    18.6     3.1000
Lack of fit        5    16.6     3.3200    1.6600    0.5272
Pure error         1     2.0     2.0000

Direction of steepest ascent (at radius 1):

     x1           x2          x3
-0.1961161   0.7844645   0.5883484

Corresponding increment in original units:

     x1           x2          x3
-0.1961161   0.7844645   0.5883484
```

We find that this output provides similar results to that obtained in MINITAB, and we may then draw the same conclusions as well.

Example 19.2.5 (Effects of time and temperature in an experiment producing an unwanted byproduct) *In a pilot plant an experimenter was interested in determining how the time and temperature conditions of a clave affected the buildup of an unwanted byproduct in a chemical process. Theoretical explanations were available, but for the purposes at hand, it was simpler merely to explore the region of interest in time and temperature by a series of experiments and to fit an approximating first-order mathematical model. To provide a measure of experimental error, the 2^2 factorial portion of the design was repeated and the* center point *replicated four times. The entire sequence of 12 runs was performed in random order. The settings of the variables time and temperature, the associated treatment matrix in the standardized variables X_1 and X_2, and the recorded responses are displayed in Table 19.2.5. The factorial points are the corners of a square.*

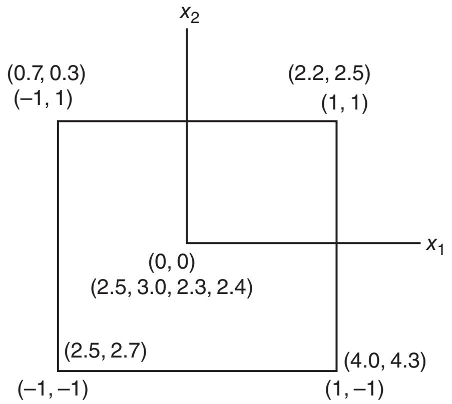


Figure 19.2.2 The 2^2 factorial design with center points in Example 19.2.5, with the observations obtained at points (runs) of this design (see Table 19.2.5).

Table 19.2.5 Treatment matrix with responses.

Time (min)	Temperature (°C)	X_1	X_2	Response Y
30	240	−1	−1	2.5
40	240	1	−1	4.0
30	250	−1	1	0.7
40	250	1	1	2.2
35	245	0	0	2.5
35	245	0	0	2.3
30	240	−1	−1	2.7
40	240	1	−1	4.3
30	250	−1	1	0.3
40	250	1	1	2.5
35	245	0	0	3.0
35	245	0	0	2.4

Solution: The first-order model is $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, and the observations y are taken to be independent $N(\eta,\ \sigma^2)$. The design employed is a 2^2 factorial with center points, replicated (see Figure 19.2.2). For this design, we have (using old and some new notation) that

$$[1] = \sum x_{i1} = \ 0, \quad [2] = \sum x_{i2} = 0, \quad [11] = \sum x_{i1}^2 = \ 8,$$
$$[22] = \sum x_{i2}^2 = \ 8, \quad [12] = \sum x_{iu}x_{iv} \ = 0$$

Thus, we can easily see that the estimates of the regression coefficients are given by

$$\hat{\beta}_0 = \frac{\sum Y_i}{N} = \frac{29.4}{12} = 2.45,\ \hat{\beta}_1 = \frac{\sum x_{i1}Y_i}{\sum x_{i1}^2} = \frac{6.8}{8} = 0.850,\ \hat{\beta}_2 = \frac{\sum x_{i2}Y_i}{\sum x_{i2}^2} = \frac{-7.8}{8} = -0.975$$

Table 19.2.6 ANOVA table for the data in Table 19.2.5.

Source	SS	DF	MS	F-ratio
X_1	5.7800	1	5.7800	88.92
X_2	7.6050	1	7.6050	117.00
Residual	0.5850	9	$0.0650 = s^2$	
Lack-of-fit	0.1050	2	0.0525	$0.765 < 1$
Pure error	0.4800	7	0.0686	
Total	13.97	11		

Table 19.2.7 Four factorial points and paired observations (see Table 19.2.5).

Treatments	(−1, −1)	(−1, 1)	(1, −1)	(1, 1)	(0, 0)
	2.5	0.7	4.0	2.2	2.5, 3.0
	2.7	0.3	4.3	2.5	2.3, 2.4
d_i	−0.2	0.4	−0.3	−0.3	$\bar{y}_c = 2.55$

so that the fitted model is

$$\hat{y} = 2.450 + 0.850x_1 - 0.975x_2 \tag{19.2.12}$$

The associated ANOVA table is displayed in Table 19.2.6.

To obtain the entry for the pure error sum of squares *SSE*, first we construct Table 19.2.7 from Table 19.2.5. For the four factorial treatments with the paired observations, using the well-known identity that $\sum_{i=1}^{2}(u_i - \bar{u})^2 = (u_1 - u_2)^2/2$, we have for their sum of squares totals $\sum_{i=1}^{4} d_i^2/2$, where d_i is the difference of the ith pair, $i = 1, 2, 3, 4$ (see Tables 19.2.7 and 19.2.5). For the center point with four observations, we compute $\sum_{i=1}^{4}(Y_i - \bar{Y}_c)^2$, where the mean $\bar{Y}_c$ of the center points is

$$\bar{Y}_c = \frac{2.5 + 2.3 + 3.0 + 2.4}{4} = 2.55$$

is the average of observations at the center point. Thus, the total pure error sum of squares is

$$[(2.5 - 2.7)^2/2 + (0.7 - 0.3)^2/2 + \cdots + (2.2 - 2.5)^2/2] \\ + [(2.5 - 2.55)^2 + \cdots + (2.4 - 2.55)^2] = 0.48$$

Note that each replicated point contributes to the pure error sum of squares. The degrees of freedom for pure error are equal to the sum of the degrees of freedom at each of the design points, which, of course, is the sum of the replications minus one at each of these points. Thus, we have degrees of freedom for pure error $= (2 - 1) + (2 - 1) + (2 - 1) + (2 - 1) + (4 - 1) = 7$.

Using the information in Table 19.2.6, a test of the hypothesis that the model is adequate to represent the unknown response function is given by the ratio of mean squares due to *lack of fit* and *pure error*, i.e., $F = 0.0525/0.0686 = 0.765$, which is not statistically significant. Thus, the model appears to be adequate. This means that we can use the residual mean square (see Table 19.2.6) to provide the estimate of the variance σ^2, which is $s^2 = 0.0650$ with $\nu = 9$ degrees of freedom. The test of the hypothesis that $\beta_1 = 0$ and that $\beta_2 = 0$ both produce very large observed $F_{1,9}$ values, so that the two hypotheses are rejected. The individual 95% confidence interval limits are, respectively,

$$\hat{\beta}_1 \pm t_{9;\ 0.025}\sqrt{S^2/[11]} = 0.850 \pm 2.26\sqrt{0.0650/8} = (0.850 \pm 0.206) = (0.644, 1.056)$$

$$\hat{\beta}_2 \pm t_{9;\ 0.025}\sqrt{S^2/[22]} = -0.975 \pm 2.26\sqrt{0.0650/8} = (-0.975 \pm 0.206) = (-1.81, -0.769)$$

The fitted equation may now be employed to map the unknown response function over the experimental region. For example, the *contour* for the predicted response $\hat{y} = 3.0$ is given by (substituting in (19.2.12)) all the points (x_1, x_2) on the line $3.000 = 2.450 + 0.850x_1 - 0.975x_2$, the equation of a straight line in the coordinate system of x_1 and x_2. The contours for $\hat{y} = 1, 2, 3$, and 4 are plotted in Figure 19.2.3 in the (X_1, X_2) coordinate system, where $x_1 = (X_1 - 35)/5,\ x_2 = (X_2 - 245)/5$ (see Table 19.2.5). These are the contours of a plane surface. Progress to a lower response (the response y is an unwanted byproduct) can be quickly explored by performing experiments along the path of steepest descent, that is, along a path perpendicular to the contour lines originating at the center of the experimental region. In Figure 19.2.3, we show a path of steepest descent originating at the center of the experimental region, which in the (X_1, X_2) coordinate system is $(X_1, X_2) = (35,\ 245)$.

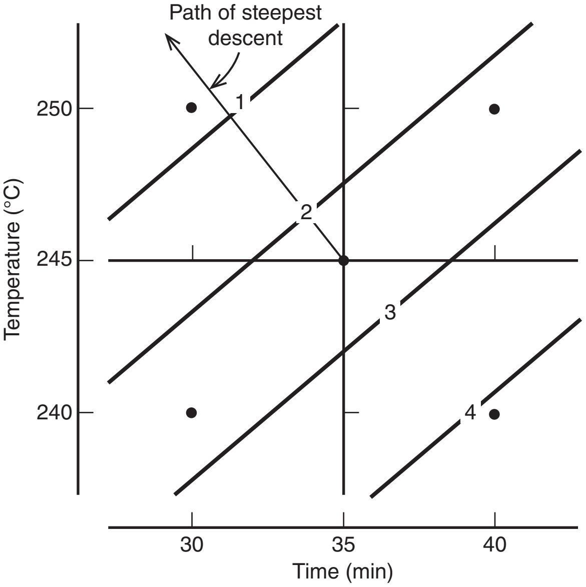


Figure 19.2.3 Contours of planar response.

PRACTICE PROBLEM FOR SECTION 19.2

1. Consider the following design plan and the data obtained using this design plan:

Run	x_1	x_2	x_3	Responses
1	−1	−1	−1	40.9
2	1	−1	−1	48.3
3	−1	1	−1	52.4
4	1	1	−1	53.7
5	−1	−1	1	55.3
6	1	−1	1	53.0
7	−1	1	1	53.4
8	1	1	1	47.0
9	0	0	0	48.9
10	0	0	0	51.7
11	0	0	0	49.7
12	0	0	0	55.0

(a) Fit a first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the above data.
(b) Find the pure error and lack of fit mean square.
(c) Prepare an ANOVA table for these data and test the adequacy of the first-order model.

2. Consider a design plan obtained by using a one-half replication $(I + ABCD)$ of a 2^4 design augmented with four center points. Suppose a fit to a first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$, using the above design plan, is made.
(a) Find the total degrees of freedom.
(b) Find the 'pure error' and 'lack of fit' degrees of freedom.
(c) Discuss whether you can test the adequacy of the first-order model. Give the test statistic for testing the fit of the first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$.

3. Suppose an engineer used the design plan given in Problem 2 in a process development study and obtained the following data.

A	B	C	D	Response
−1	−1	−1	−1	21
1	−1	−1	1	23
−1	1	−1	1	24
1	1	−1	−1	27
−1	−1	1	1	22
1	−1	1	−1	25
−1	1	1	−1	27
1	1	1	1	30
0	0	0	0	29
0	0	0	0	22
0	0	0	0	24
0	0	0	0	29

(a) Fit a first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta x_4 + \varepsilon$ to the above data.
(b) Find the 'pure error' and 'lack of fit' mean square.
(c) Prepare an ANOVA table for these data and test the adequacy of the first-order model.

4. Refer to Problem 3. Suppose because of lack of raw materials the engineer could not run the last two experiments, so the data obtained are as shown below. Repeat all the instructions of Problem 3, compare your results with those obtained earlier, and comment.

A	B	C	D	Response
−1	−1	−1	−1	21
1	−1	−1	1	23
−1	1	−1	1	24
1	1	−1	−1	27
−1	−1	1	1	22
1	−1	1	−1	25
−1	1	1	−1	27
1	1	1	1	30
0	0	0	0	29
0	0	0	0	22

5. Refer to Problem 3. Consider only the factorial points in that design and do the following:
(a) Fit a first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ to the corresponding data (i.e., the first eight lines of the table in Problem 3).
(b) Determine the alias structure for the design if the model in (a) is fitted but the true model is the second-order model $Y = \beta_0 + \Sigma\beta_i x_i + \Sigma\Sigma\beta_{ij} x_i x_j$.
(c) Estimate the regression coefficients and calculate the corresponding mean squares due to the β_i's, assuming that we are fitting the model in (a).
(d) Using the model of (a) for the factorial points, construct the ANOVA table and determine which main effects are significant. Use $\alpha = 0.05$.

6. Analyze the data of Example 19.2.5 using technology.

19.3 SECOND-ORDER DESIGNS

The general form of a second-degree polynomial in k variables is

$$y = \beta_0 + \sum_i \beta_i x_i + \sum \beta_{ii} x_i^2 + \sum \sum_{i<j} \beta_{ij} x_i x_j \tag{19.3.1}$$

In order to fit this polynomial, a total of $\dfrac{k^2 + 3k + 2}{2}$ regression coefficients are to be estimated. Further, to estimate the quadratic coefficients, each variable x_i must have at least three different levels. The designs suitable for satisfying this requirement are obviously 3^k factorial designs. For a small value of k, say $k = 2, 3$, such a design seems quite adequate. However, as k increases, the number of needed experimental points increases very rapidly. For example, when $k = 4$, in order to estimate just 15 regression coefficients, we need 81 experimental points to run a 3^4 factorial design.

Box and Wilson (1951) developed, as an alternative to the 3^k factorial designs, a class of designs called the *composite designs*. A special class of these designs is called the *central composite design*, which we study next.

19.3.1 Central Composite Designs (CCDs)

The CCDs are obtained by adding the following $(2k+n_c)$ points to a 2^k factorial or to a 2^{k-h} fractional factorial: $2k$ points are axial points and n_c points are center points. The number of center points to be chosen usually depends upon the desired degrees of freedom for the error mean square, or, the establishment of certain properties desired for the design. In general, this number is not very large. Thus, a typical *portion* of the design that is added to a 2^k factorial or fractional factorial design, as shown below, contains $2k$ axial points and n_c center points, illustrated below.

$$\left[\begin{array}{cccc}
x_1 & x_2 & \cdots & x_k \\
-\alpha & 0 & \cdots & 0 \\
\alpha & 0 & \cdots & 0 \\
0 & -\alpha & \cdots & 0 \\
0 & \alpha & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & -\alpha \\
0 & 0 & \cdots & \alpha \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & 0
\end{array}\right]$$

Thus, if we use a complete replication of a 2^k factorial design, the total number of points in this CCD design is $(2^k + 2k + n_c)$. For $k = 2, 3, 4$, we have $8 + n_c$, $14 + n_c$, and $24 + n_c$ points, respectively, whereas using 3^k factorials would lead to CCDs for $k = 2, 3, 4$ consisting of 9, 27, and 81 points, respectively. Thus, for k as small as 4 in 3^k factorials, we have to perform almost three times the number of experiments as in the CCDs. The value of α (where $|\alpha| > 1$) depends on certain properties desired for the design and on the number of factors involved. For two independent variables x_1, x_2 the CCD, when viewed geometrically, has the 2^2 vertices of a square, $k = 2$ points on each axis, n_c points at the center (origin), as illustrated in Figure 19.3.1.

Note that experimental runs at the *axial points* are identical to the center points *except for one* factor. In other words, the factors are varying not simultaneously but one at a time. As a result the observations at these points provide no information on the interaction term but allow for estimation of the pure quadratic terms. The center points provide $(n_c - 1)$ degrees of freedom for the pure error mean square and allow the testing of various hypotheses. They also provide one degree of freedom for estimating the sum of the pure quadratic terms.

As mentioned above, the value of α to be chosen depends on certain properties desired for the design and on the number of factors involved. One of the desirable properties of a

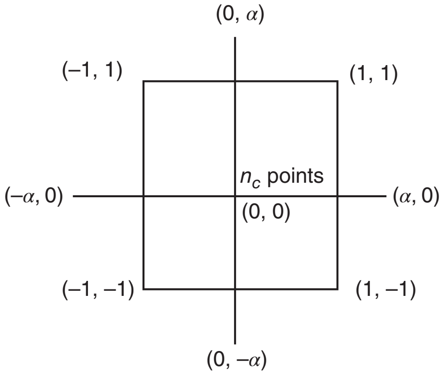


Figure 19.3.1 Geometry of a CCD that uses a 2^2 replication with n_c center points.

CCD is *rotatability*, a concept introduced by Box and Hunter (1957). A *rotatable design* allows $\hat{y}$ at points $x = (x_1, \dots, x_k)$, say $\hat{y}(x)$, to be such that $Var(\hat{y}(x))$ is the same at all points that are at the same distance from the center of the design. In other words, the precision of the predicted response is the same at all points on a sphere in k dimensions. A CCD that possesses the property of rotatability depends upon the value of α and the number of experimental runs in the factorial portion of the CCD (see Myers and Montgomery 1995). If the number of experimental runs in the factorial portion is F, the value of α needed to make the design rotatable is given by

$$\alpha = \sqrt[4]{F} \tag{19.3.2}$$

Table 19.3.1 provides some typical values of α as a function of the number of experimental runs in the factorial portion of the CCDs (see Myers and Montgomery 1995).

Table 19.3.1 Values of α to establish a rotatable central composite design.

Design	2^2	2^3	2^4	2^5	2^{5-1}	2^6	2^{6-1}	2^7	2^{7-1}
F	4	8	16	32	16	64	32	128	64
α	1.414	1.682	2.000	2.378	2.000	2.828	2.378	3.364	2.828

The major advantage of the CCDs is that they can be performed in different stages. The first-order design, which corresponds to a 2^k factorial or a fractional factorial plus some center points, is run in a first stage. This design is used to fit a first-degree polynomial and to look for any indication of curvature. If the experimenter finds the *lack of fit* term in the analysis is significant, then in the second stage, experiments are run at the $2k$ axial points. The second stage experiments allow the estimation of the regression coefficients of second-degree terms. Thus, combining the results of the experiments performed in two stages, we can fit a second-degree polynomial and test for lack of fit.

If an experiment that uses a CCD requires blocking, then a CCD allows much flexibility in establishing such a design. In fact, under certain conditions, the experiment can be carried out in orthogonal blocks.

Definition 19.3.1 Blocks or factors are called orthogonal if (i) $\sum_{i=1}^{n_j} x_{iu}x_{iw} = 0$, where $u \neq w = 0, 1, \cdots, k$, and where x_{iu} and x_{iw} are the ith entries of the uth and wth factors, with $x_{i0} = 1$, for all i. Further, n_j is the number of treatments in the jth block, $j = 1, ..., m$, where m is the number of blocks and k is the number of factors in each block. (ii) Orthogonality also requires that $\sum_{i=1}^{n_j} x_{iu}^2 / \sum_{w=1}^{k} \sum_{i=1}^{N} x_{iw}^2 = n_j/N$, where N = total number of entries of each factor in all the blocks.

For a detailed discussion on orthogonal blocking in CCD designs, see Myers and Montgomery (1995), Khuri and Cornell (1996), and Box and Draper (1987).

Example 19.3.1 (Experiment for achieving high concentration of a chemical) *Two chemicals, A and B, are combined and as the reaction takes place, a new chemical C forms. The concentration of C depends upon the concentrations of A and B and the temperature maintained at which the reaction takes place. The aim of the experiment is to get the highest concentration of C. The three factors varied are the temperature T, the concentration of A, and the concentration of B. The first set of experiments is performed by taking two levels of each factor as shown below*

Factors	levels
t (temperature,°C)	130, 150
a (% of A)	25, 30
b (% of B)	35, 40

Solution: Let the variables t, a, and b, when expressed in standard units of factorial design, be denoted by x_1, x_2, and x_3, respectively. Then the relation between the variables x_1, x_2, and x_3 and the natural variables t, a, and b are taken as follows:

$$x_1 = \frac{t - 140}{10}, \quad x_2 = \frac{a - 27.5}{2.5}, \quad x_3 = \frac{b - 37.5}{2.5}$$

Thus, the levels of each factor, when expressed in standard units, are (−1, 1). The trials were carried out randomly. The yields are shown in Table 19.4.2.

Table 19.3.2 Percentage of concentration of C.

Treatment	x_1	x_2	x_3	% C
1	−1	−1	−1	18.2
2	1	−1	−1	22.8
3	−1	1	−1	22.8
4	1	1	−1	23.0
5	−1	−1	1	22.7
6	1	−1	1	23.2
7	−1	1	1	24.2
8	1	1	1	20.8

Suppose now the experimenter fits the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \tag{19.3.3}$$

but the true model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \tag{19.3.4}$$

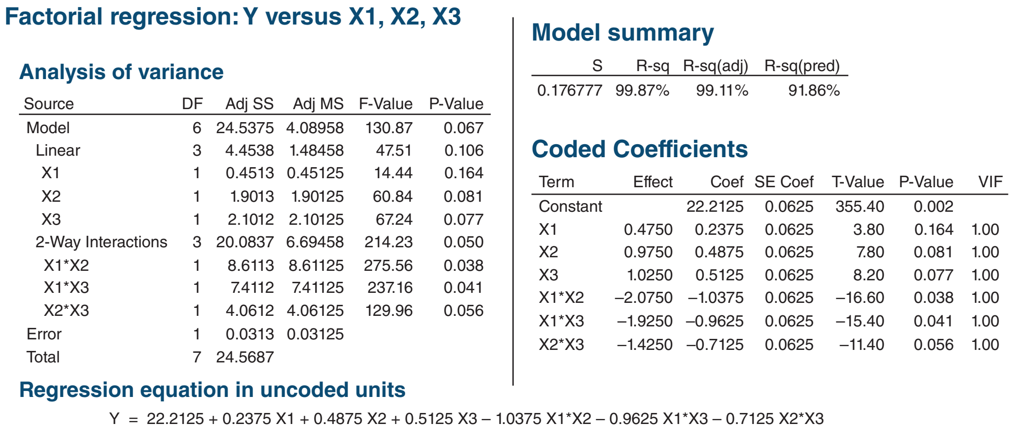
The experimenter is, of course, fitting model (19.3.3) to the data in Table 19.3.2. The MINITAB printout shows that the fit to the data set in Table 19.3.2 of the model (19.3.3) is adequate (R-Sq(adj) = 99.11%). Now if (19.3.4) is the true model, it turns out that $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ are unbiased for β_1, β_2, β_3 respectively, but that $\hat{\beta}_0$ is a biased estimator in that

$$E(\hat{\beta}_0) = \beta_0 + \beta_{11} + \beta_{22} + \beta_{33} \tag{19.3.4a}$$

In fact, for the data set in Table 19.3.2, we find that the estimates of the linear regression coefficients and the two way interaction coefficients are, respectively.

$\hat{\beta}_0 = 22.21$, $\hat{\beta}_1 = 0.238$, $\hat{\beta}_2 = 0.488$, $\hat{\beta}_3 = 0.512$, $\hat{\beta}_{12} = -1.037$, $\hat{\beta}_{13} = -0.962$, and $\hat{\beta}_{23} = -0.712$

To run the model specify in (19.3.3), we proceed in MINITAB as in Example 19.2.4, but in Step 7, we proceed as: From the bar menu select **Stat** > **DOE** > **Response Surface** > **Analyze Response Surface Design.** Then enter C5 under the box Responses: this appears in the new window and click **Terms** and select **Linear + interactions** when the pull down menu appears. Click **OK** twice. The MINITAB output appears in the **Session Window** as shown below.



Factorial regression: Y versus X1, X2, X3

Analysis of variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	6	24.5375	4.08958	130.87	0.067
Linear	3	4.4538	1.48458	47.51	0.106
X1	1	0.4513	0.45125	14.44	0.164
X2	1	1.9013	1.90125	60.84	0.081
X3	1	2.1012	2.10125	67.24	0.077
2-Way Interactions	3	20.0837	6.69458	214.23	0.050
X1*X2	1	8.6113	8.61125	275.56	0.038
X1*X3	1	7.4112	7.41125	237.16	0.041
X2*X3	1	4.0612	4.06125	129.96	0.056
Error	1	0.0313	0.03125		
Total	7	24.5687			

Model summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.176777	99.87%	99.11%	91.86%

Coded Coefficients

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		22.2125	0.0625	355.40	0.002	
X1	0.4750	0.2375	0.0625	3.80	0.164	1.00
X2	0.9750	0.4875	0.0625	7.80	0.081	1.00
X3	1.0250	0.5125	0.0625	8.20	0.077	1.00
X1*X2	−2.0750	−1.0375	0.0625	−16.60	0.038	1.00
X1*X3	−1.9250	−0.9625	0.0625	−15.40	0.041	1.00
X2*X3	−1.4250	−0.7125	0.0625	−11.40	0.056	1.00

Regression equation in uncoded units

```
Y = 22.2125 + 0.2375 X1 + 0.4875 X2 + 0.5125 X3 − 1.0375 X1*X2 − 0.9625 X1*X3 − 0.7125 X2*X3
```

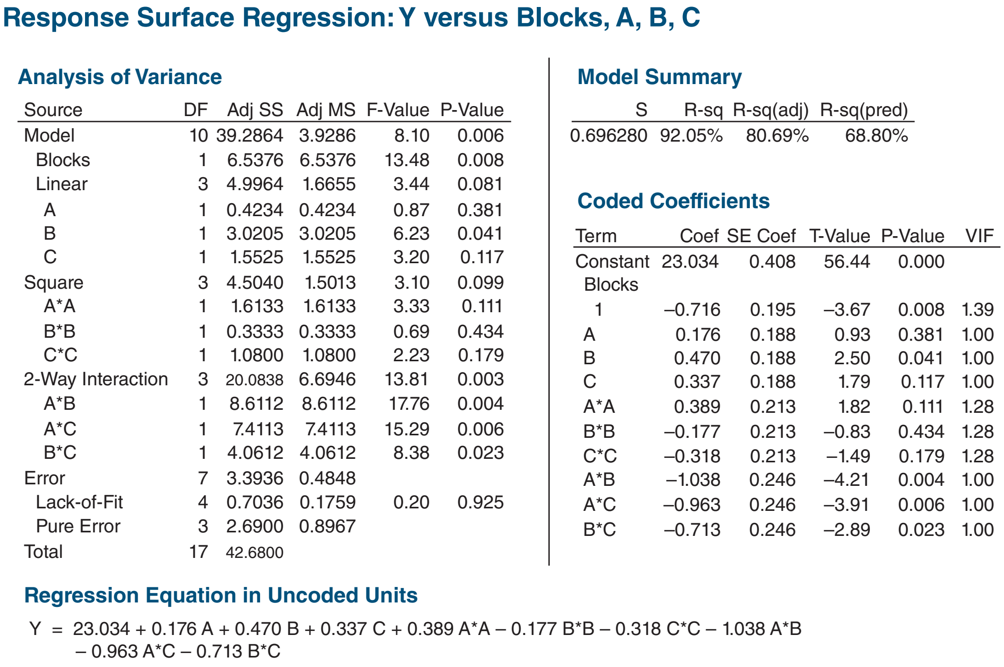
We note that two of the two-factor interaction coefficients are significant at the 5% level, while the third two-factor interaction is significant at the 5.6% level, i.e. barely missing the 5% level. This indicates that the true response surface may have some curvature. Note that if the interactions are not significant, then it is still possible that the quadratic terms will be present in the fitted model. We now remark that it is always better to add some center points to the 2^3 design, as this gives us some degrees of freedom for the pure error along with one degree of freedom for the sum of the coefficients of the pure quadratic terms. This may then be utilized to help decide whether there is any curvature in the data. The significance of the interaction and quadratic terms usually indicates that we are near

Table 19.3.3 Ten extra points and their observed percentage of concentration of C.

Treatment	x_1	x_2	x_3	$\%\ C$
9	0	0	0	23.6
10	0	0	0	24.9
11	0	0	0	22.6
12	0	0	0	23.9
13	0	0	-1.682	22.7
14	0	0	1.682	23.0
15	-1.682	0	0	24.7
16	1.682	0	0	25.0
17	0	-1.682	0	22.5
18	0	1.682	0	24.0

the optimum point. Thus, in this case, it is worth fitting the second-degree polynomial so that we now consider a new set of 10 points (6 axial points and 4 center points). Combining these points with the previous eight points, we have a second-order CCD. The values of α $(= 8^{1/4})$ are selected so that the CCD is rotatable or near-rotatable.

The MINITAB output appears in the **Session Window** for the model after adding above 10 points as shown below (Note: The MINITAB step by step procedure is discussed subsequently).



```
Response Surface Regression: Y versus Blocks, A, B, C

Analysis of Variance

Source             DF   Adj SS   Adj MS  F-Value  P-Value
Model              10  39.2864   3.9286     8.10    0.006
  Blocks            1   6.5376   6.5376    13.48    0.008
  Linear            3   4.9964   1.6655     3.44    0.081
    A               1   0.4234   0.4234     0.87    0.381
    B               1   3.0205   3.0205     6.23    0.041
    C               1   1.5525   1.5525     3.20    0.117
Square              3   4.5040   1.5013     3.10    0.099
    A*A             1   1.6133   1.6133     3.33    0.111
    B*B             1   0.3333   0.3333     0.69    0.434
    C*C             1   1.0800   1.0800     2.23    0.179
2-Way Interaction   3  20.0838   6.6946    13.81    0.003
    A*B             1   8.6112   8.6112    17.76    0.004
    A*C             1   7.4113   7.4113    15.29    0.006
    B*C             1   4.0612   4.0612     8.38    0.023
Error               7   3.3936   0.4848
  Lack-of-Fit       4   0.7036   0.1759     0.20    0.925
  Pure Error        3   2.6900   0.8967
Total              17  42.6800

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.696280  92.05%     80.69%      68.80%

Coded Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  23.034    0.408    56.44    0.000
 Blocks
  1       -0.716    0.195    -3.67    0.008  1.39
 A         0.176    0.188     0.93    0.381  1.00
 B         0.470    0.188     2.50    0.041  1.00
 C         0.337    0.188     1.79    0.117  1.00
 A*A       0.389    0.213     1.82    0.111  1.28
 B*B      -0.177    0.213    -0.83    0.434  1.28
 C*C      -0.318    0.213    -1.49    0.179  1.28
 A*B      -1.038    0.246    -4.21    0.004  1.00
 A*C      -0.963    0.246    -3.91    0.006  1.00
 B*C      -0.713    0.246    -2.89    0.023  1.00

Regression Equation in Uncoded Units

Y  =  23.034 + 0.176 A + 0.470 B + 0.337 C + 0.389 A*A – 0.177 B*B – 0.318 C*C – 1.038 A*B
      – 0.963 A*C – 0.713 B*C
```

The above analysis shows that a complete second-degree polynomial (interaction terms $x_i x_j$ and pure quadratic terms are in this model) is a very good fit. Indeed, the observed no-lack-of-fit test statistic is 0.20, and obviously not significant. This indicates we may be near to the optimal point. But before doing further experiments, we first analyze the fitted response surface (we do this later in Section 19.4). The fitted response surface in this example, based on the design and observations found in Tables 19.3.2 and 19.3.3 is given by

$$\hat{Y} = 23.0343 + 0.1761x_1 + 0.4703x_2 + 0.3372x_3 + 0.3889x_1^2 - 0.1768x_2^2 - 0.3182x_3^2 - 1.0375x_1x_2 - 0.9525x_1x_3 - 0.7125x_2x_3 \quad (19.3.5)$$

Figures 19.3.2 and 19.3.3 show the contour plots in the (x_1, x_2) coordinate system and response surface plots for the fitted response surface model (19.3.5). The response surface indicates that the stationary point is a saddle point. These figures give in the box at top right, the value of the third factor as the other two factors vary. Note that the contour plots can be used to develop the operating conditions (if they exist within the experimental region) on the design variables under which the optimum value of the response variable can be achieved. It is not a good idea to define any operating conditions outside the experimental region, since the fitted model is not reliable for conditions outside that region.

We now show all the steps needed to fit a response surface using MINITAB and R.

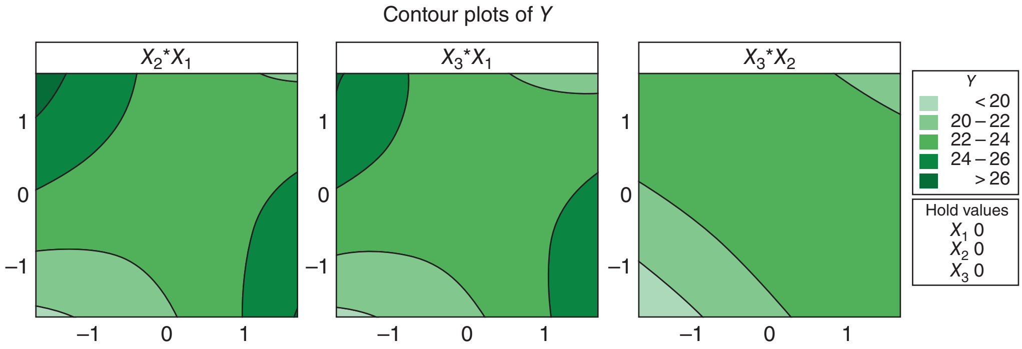


Figure 19.3.2 Contour plots for the fitted response surface model (19.3.5).

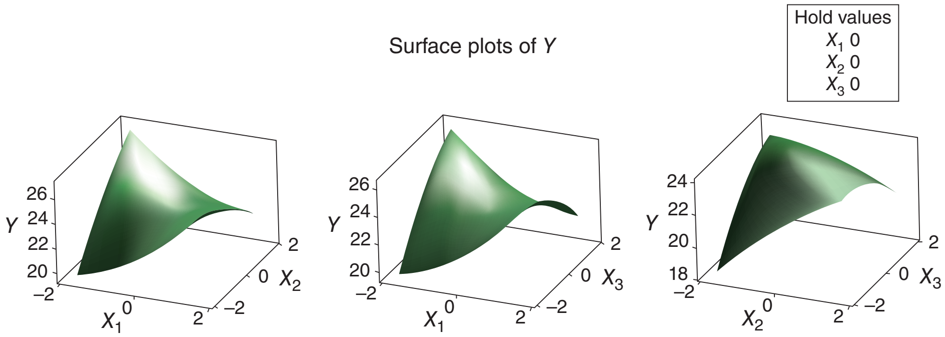


Figure 19.3.3 Response surface plots for the fitted response surface model (19.3.5).

MINITAB

If the experimenter has performed the experiment in two stages (see Example 19.4.1), then we proceed as follows:

1. From the bar menu select **Stat** > **DOE** > **Factorial** > **Create Factorial Design . . . .**
2. Check **2-level factorial** (default generators), select the number of factors, and click on **Designs . . .** option.
3. Select full replication, then select (i) number of center points per block (0), (ii) number of replicates for corner points (1), and (iii) number of blocks (1), and then click **OK** twice.
4. From the Menu bar select **Stat** > **DOE** > **Modify Design ...**
5. Select**Add axial points** and then click **Specify.**In the new dialog box, select the value of alpha (e.g., 1.681793) and the number of center points and then click **OK**. The desired second-order design will appear in the **Worksheet.**
6. Enter the data in the next empty column of the Worksheet and proceed as follows to analyze the data.
7. From the bar menu, select **Stat** > **DOE** > **Response Surface** > **Analyze Response Surface Design . . . .** In the new dialog box "**Analyze Response Surface Design**," enter the column of responses under **Response**: and select **Terms . . .** and select Full quadratic from the pull down menu. If there is any particular quadratic term(s), you do not want to include in the model, click on that term twice (this moves these terms to the box under available terms). Then click **OK** twice. The complete analysis of the data appears in the session window (see Example 19.3.1).
8. For contour plots and surface plots, select from the bar menu **Stat** > **DOE** > **Response Surface** > **Contour plot** or **Surface plots**. Then in the "**Contour**" or "**Surface Plots**" dialog box that appears select the appropriate choices for the graphical displays you desired.

If the experiment was performed in one stage, then we can still create the design as above or we can take the following steps:

1. From the bar menu select **Stat** > **DOE** > **Response Surface** > **Create Response Surface Design.**
2. Check **Central composite** and select the number of factors and click on **Designs . . .** option.
3. Select either of the available designs and check Number of Center points and value of alpha by default, or select custom and enter, in the box next to the Cube block, the number of center points. Then follow steps 6–8.
4. From the Menu bar select **Stat** > **DOE** > **Modify Design . . . .**

USING R

Solution To perform the required response surface modeling in R, we can use the 'rsm()' function in the R 'library(rsm)' as shown in the following R-code. Both 'FO' and ‘TWI’ terms are included as both first-order and two-way interaction terms are required. Additional functions ‘contour()’ and ‘persp()’ are used to get required contour and surface plots.

```
library(rsm)
#Full factorial model
x1 = c(-1,1,-1,1,-1,1,-1,1)
x2 = c(-1,-1,1,1,-1,-1,1,1)
x3 = c(-1,-1,-1,-1,1,1,1,1)
y = c(18.2,22.8,22.8,23,22.7,23.2,24.2,20.8)
data = data.frame(x1, x2, x3, y)

mod.rsm = rsm(y ~ FO(x1, x2, x3) + TWI(x1, x2, x3), data = data)
summary(mod.rsm)

#Full factorial with center and axial points
x12 = c(-1,1,-1,1,-1,1,-1,1,0,0,0,0,0,0,-1.682,1.682,0,0)
x22 = c(-1,-1,1,1,-1,-1,1,1,0,0,0,0,0,0,0,0,-1.682,1.682)
x32 = c(-1,-1,-1,-1,1,1,1,1,0,0,0,0,-1.682,1.682,0,0,0,0)
Blocks = c(-1,-1,-1,-1,-1,-1,-1,-1,1,1,1,1,1,1,1,1,1,1)
y2 = c(18.2,22.8,22.8,23,22.7,23.2,24.2,20.8,23.6,24.9,22.6,23.9,
22.7,23,24.7,25,22.5,24)
data2 = data.frame(x12, x22, x32, Blocks, y2)

mod.rsm = rsm(y2 ~ Blocks + SO(x12, x22, x32) + TWI(x12, x22, x32), data = data2)
summary(mod.rsm)

par(mfrow = c(2, 3))
contour(mod.rsm, ~ x12 + x22 + x32, at=c(Blocks=-1, x12=0, x22=0, x32=0), image =
TRUE)
persp(mod.rsm, ~ x12 + x22 + x32, zlab=“y,” at=c(Blocks=-1, x12=0, x22=0, x32=0), col
= “green,” theta = 135, phi = 20)
```

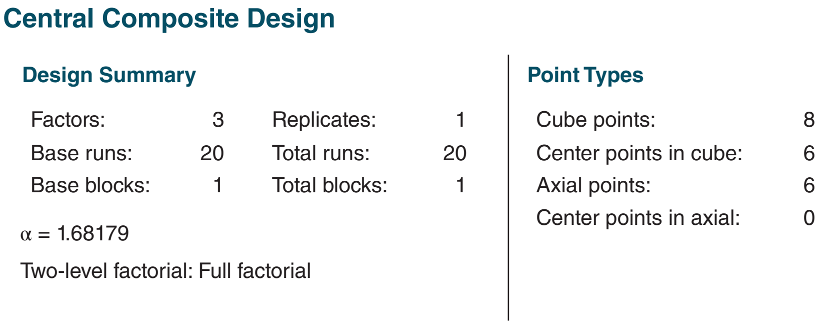
We find that this output provides similar results to that obtained in MINITAB, and we may then draw the same conclusions as well.

Example 19.3.2 (Experiment for achieving high yield of a chemical) *In an attempt to increase production, a chemical engineer studied three factors, A, B (temperature), and C (pH). For the first stage, it was proposed to use each factor at two levels and to add four center points. However, after the analysis of these data, further investigation was proposed. Thus, eight extra points (six axial and two center points) were added in such a way that the design used was a central composite rotatable design. Table 19.3.4 gives the complete plan and the data on the yield. Fit a second-degree polynomial to the data in Table 19.3.4.*

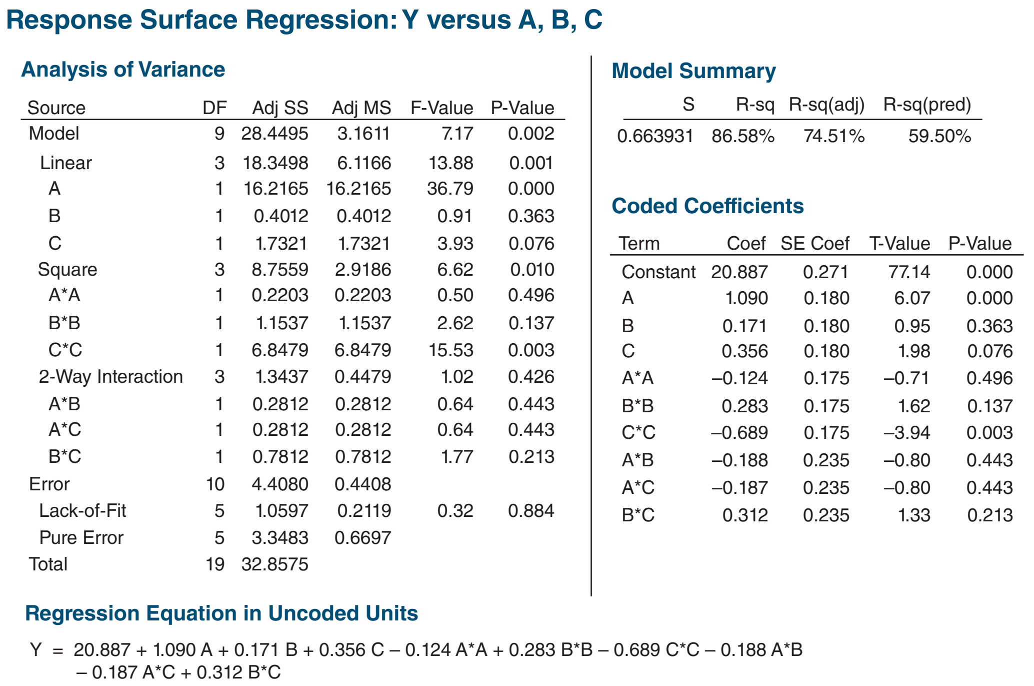
Table 19.3.4 Data obtained using a central composite rotatable design.

−1	−1	−1	18.5
1	−1	−1	22.0
−1	1	−1	19.0
1	1	−1	21.0
−1	−1	1	19.0
1	−1	1	21.0
−1	1	1	20.0
1	1	1	22.0
0	0	0	21.7
0	0	0	20.2
0	0	0	20.5
0	0	0	21.6
0	0	0	19.8
0	0	0	21.5
−1.682	0	0	19.0
1.682	0	0	22.2
0	−1.682	0	21.5
0	1.682	0	22.0
0	0	−1.682	18.0
0	0	1.682	20.0

To run this design in MINITAB, we proceed as follows: From the bar menu select **Stat** > **DOE** > **Response Surface** > **Create Response Surface Design.** Then select **Central composite** option from the new window appears. Then select the appropriate number of continuous factors and click on **Display Available Designs...** and select **central composite full** design with 3 **unblocked** factors which results in 20 test runs from the table that appears in the new window. Click **OK** and check other options if needed. Click **OK**. The desired CCD will appear in the **Worksheet**. Enter the data in the next empty column of the Worksheet and proceed as follows to analyze the data. For analysis, we follow the MINITAB steps 7–8 in Example 19.3.1. The following MINITAB output and contour plots in Figure 19.3.4 and surface plots in Figure 19.3.5 will be appeared.



```
Central Composite Design

Design Summary                                        Point Types

Factors:          3    Replicates:      1            Cube points:              8
Base runs:       20    Total runs:     20            Center points in cube:    6
Base blocks:      1    Total blocks:    1            Axial points:             6
                                                     Center points in axial:   0
α = 1.68179

Two-level factorial: Full factorial
```



```
Response Surface Regression: Y versus A, B, C

Analysis of Variance                                          Model Summary

Source              DF   Adj SS   Adj MS  F-Value  P-Value           S     R-sq  R-sq(adj)  R-sq(pred)
Model                9  28.4495   3.1611     7.17    0.002    0.663931   86.58%     74.51%      59.50%
  Linear             3  18.3498   6.1166    13.88    0.001
    A                1  16.2165  16.2165    36.79    0.000
    B                1   0.4012   0.4012     0.91    0.363    Coded Coefficients
    C                1   1.7321   1.7321     3.93    0.076
  Square             3   8.7559   2.9186     6.62    0.010    Term        Coef  SE Coef  T-Value  P-Value
    A*A              1   0.2203   0.2203     0.50    0.496    Constant  20.887    0.271    77.14    0.000
    B*B              1   1.1537   1.1537     2.62    0.137    A          1.090    0.180     6.07    0.000
    C*C              1   6.8479   6.8479    15.53    0.003    B          0.171    0.180     0.95    0.363
  2-Way Interaction  3   1.3437   0.4479     1.02    0.426    C          0.356    0.180     1.98    0.076
    A*B              1   0.2812   0.2812     0.64    0.443    A*A       –0.124    0.175    –0.71    0.496
    A*C              1   0.2812   0.2812     0.64    0.443    B*B        0.283    0.175     1.62    0.137
    B*C              1   0.7812   0.7812     1.77    0.213    C*C       –0.689    0.175    –3.94    0.003
Error               10   4.4080   0.4408                      A*B       –0.188    0.235    –0.80    0.443
  Lack-of-Fit        5   1.0597   0.2119     0.32    0.884    A*C       –0.187    0.235    –0.80    0.443
  Pure Error         5   3.3483   0.6697                      B*C        0.312    0.235     1.33    0.213
Total               19  32.8575

Regression Equation in Uncoded Units

Y = 20.887 + 1.090 A + 0.171 B + 0.356 C – 0.124 A*A + 0.283 B*B – 0.689 C*C – 0.188 A*B
    – 0.187 A*C + 0.312 B*C
```

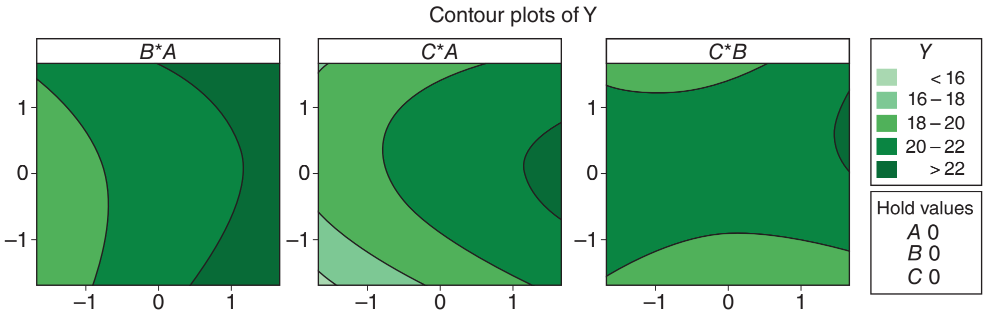


Figure 19.3.4 Contour plots for the fitted response surface of Example 19.3.2.

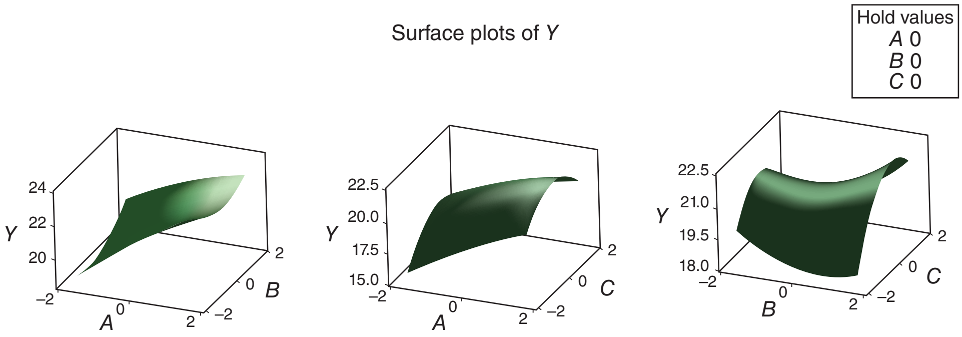
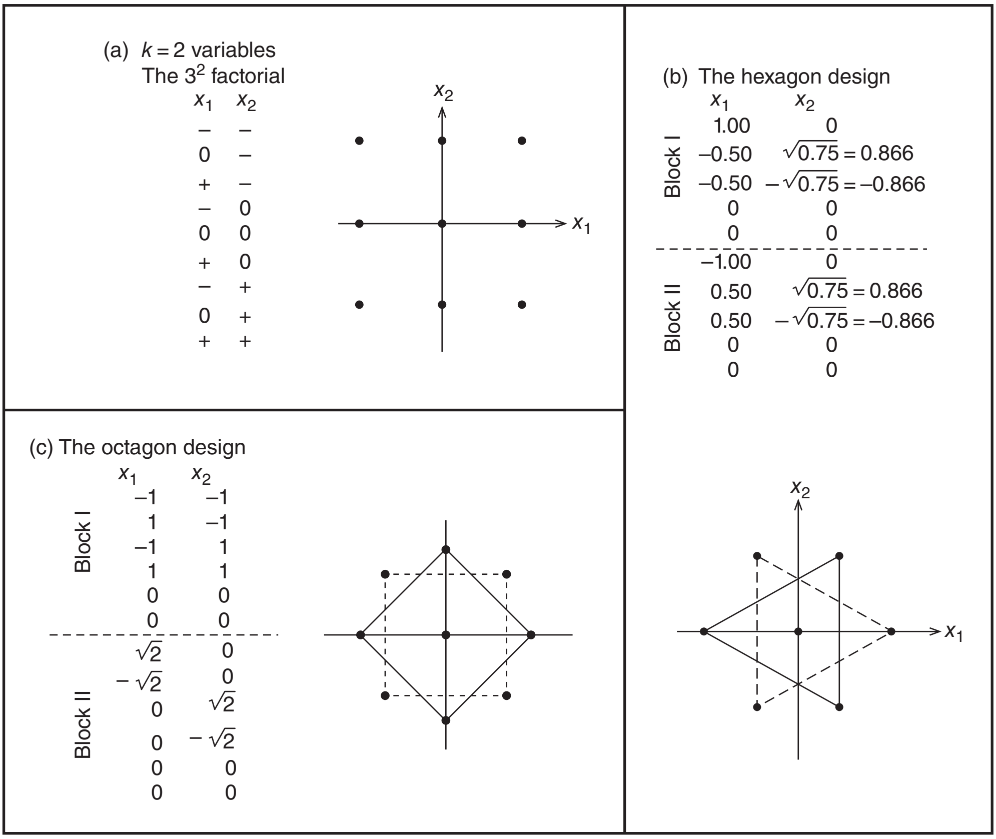


Figure 19.3.5 Response surface plots for the fitted response surface of Example 19.3.2.

Note that the box on the top right of these figures gives the value of the third factor as the other two factors vary.

19.3.2 Some Other First-Order and Second-Order Designs

Some other second-order designs are useful in applications. We give a few of them in Figures 19.3.6 and 19.3.7.



(a) $k = 2$ variables
The 3^2 factorial

x_1	x_2
−	−
0	−
+	−
−	0
0	0
+	0
−	+
0	+
+	+

(b) The hexagon design

	x_1	x_2
Block I	1.00	0
	−0.50	$\sqrt{0.75} = 0.866$
	−0.50	$-\sqrt{0.75} = -0.866$
	0	0
	0	0
Block II	−1.00	0
	0.50	$\sqrt{0.75} = 0.866$
	0.50	$-\sqrt{0.75} = -0.866$
	0	0
	0	0

(c) The octagon design

	x_1	x_2
Block I	−1	−1
	1	−1
	−1	1
	1	1
	0	0
	0	0
Block II	$\sqrt{2}$	0
	$-\sqrt{2}$	0
	0	$\sqrt{2}$
	0	$-\sqrt{2}$
	0	0
	0	0

Figure 19.3.6 Some second-order designs.

Example 19.3.3 (Data on crystal study) *In a study to determine the optimum conditions for the growth of large ZnS crystals of great purity, two factors were varied: the temperature of the melt and the rate of withdrawal of the crucible in which the crystal was grown. The experimenters began their investigation by employing a replicated simplex design (geometrically, a simplex design for $k = 2$, are the vertices of an equilateral triangle) with repeated center points. The factor settings, settings of the experimental design variables, and a recorded response (here coded) are displayed in Table 19.3.5. All treatments were run in random order.*

The central composite design

	x_1	x_2	x_3
Block I	−1	−1	1
	1	−1	−1
	−1	1	−1
	1	1	1
	0	0	0
	0	0	0
Block II	−1	−1	−1
	1	−1	1
	−1	1	1
	1	1	−1
	0	0	0
	0	0	0
Block III	1.63	0	0
	−1.63	0	0
	0	−1.63	0
	0	1.63	0
	0	0	−1.63
	0	0	1.63
	0	0	0
	0	0	0

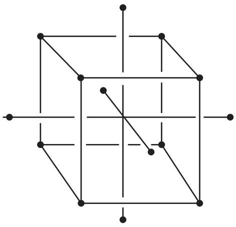


Figure 19.3.7 Central composite design with orthogonal blocking.

Table 19.3.5 The settings of the experimental design variables and recorded responses.

Factor settings						
Temperature (°C)	Rate	Design variables		Response y		Computations
	(in./d)	x_1	x_2			
1920	1.00	1.00	0	7.2	6.9	$N = 10, \Sigma y = 101.5$
1890	1.05	−0.50	0.866	9.3	9.6	$\Sigma y^2 = 1065.73$
1890	0.95	−0.50	−0.866	10.4	9.8	$\Sigma x_1 y = -5.4500$
1900	1.00	0	0	12.3	11.7	$\Sigma x_2 y = -1.1258$
1900	1.00	0	0	12.2	12.1	$[11] = [22] = 3.0$

To illustrate some of the above computations, we have the following: There are two observations per treatment, so we have

$$\begin{aligned}\sum x_1 y &= (1)(7.2) + (1)(6.9) + (-0.5)(9.3) + (-0.5)(9.6) + (-0.5)(10.4) + (-0.5)(9.8)\\ &\quad + 0(12.3) + 0(11.7) + 0(12.2) + 0(12.1) = -5.45\\ [11] = \sum x_i^2 &= 1^2 + 1^2 + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 + 0^2 + 0^2 + 0^2 + 0^2\\ &= 3.00\end{aligned}$$

Solution: The object of the above experimental program was to see if a plane would be an acceptable approximation to the response function and if a *path of steepest ascent* (see Section 19.4) determined in the space of the factors would lead to a region of higher responses. The data in Table 19.3.5 are plotted in Figure 19.3.8 and it is obvious that a plane is inadequate to represent the response function. The observed response at the center of the experimental region appears to be larger than the average response at the external

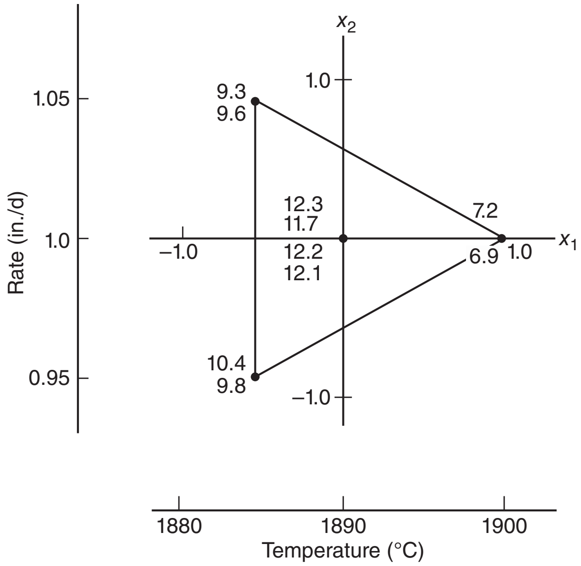


Figure 19.3.8 Simplex design for fitting a first-order model.

design points, an indication that the response surface has a curvature. The hypothesis that the surface is nonplanar can be verified by fitting the first-order model and then testing to see whether it is adequate to represent the data. Fitting the model $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ to these data gives

$$\hat{\beta}_0 = \frac{\sum y}{N} = \frac{101.5}{10} = 10.15, \quad \hat{\beta}_1 = \frac{\sum x_{i1} y}{\sum x_{i1}^2} = \frac{-5.45}{3.0} = -1.8167,$$

$$\hat{\beta}_2 = \frac{\sum x_{i2} y}{\sum x_{i2}^2} = \frac{-1.1258}{3.0} = -0.3753$$

so that the fitted model is

$$\hat{y} = 10.15 - 1.82 x_1 - 0.38 x_2$$

The corresponding analysis of variance is displayed in Table 19.3.6.

Table 19.3.6 ANOVA table for first-order design.

Source	*DF*	*SS*	*MS*	*F*-ratio
Linear	2	10.3235	5.1617	
Residual	7	25.1815		
Lack of fit	1	24.7040	24.7040	310.35
Pure error	6	0.4775	0.0796 $=S^2$	
Total	9	35.5050		

Now under the H_0 : no lack of fit, the ratio of the lack of fit mean square to the error mean square is distributed as a $F_{1,6}$ random variable, and in any case, it is obvious that the observed ratio $F = 24.7040/0.0796 = 310.35$ is highly unusual.

An alternative, and for this design, the exactly equivalent test of the hypothesis that no curvature exists, is provided by comparing the observations at the center of the design against those of the periphery or exterior points by constructing a relevant contrast.

Now recall from Chapter 18 that a contrast is a linear combination of observations, with the constants in the linear combination summing to zero. Denote the observations taken at (0, 0), (1, 0), (−0.5, 0.866), (−0.5, −0.866) by y_{0j} $(j = 1, 2, 3, 4)$; y_{1j} $(j = 1, 2)$; y_{2j} $(j = 1, 2)$; y_{3j} $(j = 1, 2)$, respectively. To compare the observations taken at (0, 0), the center of the design, with those taken at the exterior points, namely at (1, 0), (−0.5, 0.866), (−0.5, −0.866), we use the contrast

$$3(y_{01} + y_{02} + y_{03} + y_{04}) - 2(y_{11} + y_{12}) - 2(y_{21} + y_{22}) - 2(y_{31} + y_{32})$$

(Note that the constants in this contrast sum to zero.) Alternatively, we may write this as

$$3(4)\bar{y}_0 - 2(2)\bar{y}_1 - 2(2)\bar{y}_2 - 2(2)\bar{y}_3$$

which is of the form $\sum_{j=0}^{3} c_j d_j \bar{y}_j$, with d_j the number of observations at each treatment. That is,

$$c_0 = 3, c_1 = -2, c_2 = -2, c_3 = -2$$

$$d_0 = 4, d_1 = 2, d_2 = 2, d_3 = 2$$

and we note again that $\sum_{j=0}^{3} c_j d_j = 0$. Inserting the actual observations from Table 19.3.5 or the information below, we find that $\sum_{j=0}^{3} c_j d_j \bar{y}_j = 38.50$.

The corresponding single degree of freedom sum of squares is (see Section 18.1)

$$\frac{\left(\sum c_j (d_j \bar{y}_j)\right)^2}{\left(\sum d_j c_j^2\right)} = \frac{(38.50)^2}{4(9) + 2(4) + 2(4) + 2(4)} = \frac{(38.50)^2}{60} = 24.7042$$

which, except for the rounding errors, equals the lack of fit sum of squares in Table 19.3.6. The hypothesis that the contrast effect is zero, that is, that no curvature exists, is thus rejected with the observed lack of fit $F = 24.7042/0.0796 = 310$, obviously significantly large.

After reviewing these data, the experimenters decided to form a hexagon design by adding a second replicated simplex design with four center points as illustrated in Figure 19.3.9. The factor settings, level of the design variables, and observed responses are displayed in Table 19.3.7.

The proposed second-order model is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

The corresponding normal equations are (see Section 19.5)

$$\begin{aligned}
20\hat{\beta}_0 \qquad\qquad\qquad & + 6.0\hat{\beta}_{11} + 6.0\hat{\beta}_{22} = 200.8000 \\
6.0\hat{\beta}_1 \qquad\qquad & \qquad\qquad\qquad\quad = -3.5000 \\
6.0\hat{\beta}_2 \qquad & \qquad\qquad\qquad\quad = -10.5652 \\
6.0\hat{\beta}_0 \qquad\qquad\qquad & + 4.5\hat{\beta}_{11} + 1.5\hat{\beta}_{22} = 48.1000 \\
6.0\hat{\beta}_0 \qquad\qquad\qquad & + 1.5\hat{\beta}_{11} + 4.5\hat{\beta}_{22} = 55.5000 \\
& \qquad\qquad 1.5\hat{\beta}_{12} = -4.1568
\end{aligned} \qquad (19.3.6)$$

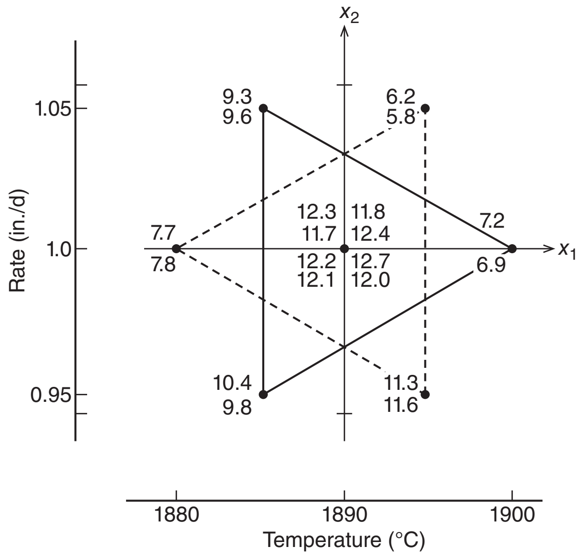


Figure 19.3.9 Hexagon Design in two blocks for fitting a second-order model.

Table 19.3.7 The factor setting, levels of the design variables, and observed responses.

Factor settings					
Temperature (°C)	Rate	Design variables		Response y	
	(in./d)	x_1	x_2		
1920	1.00	1.00	0	7.2	6.9
1890	1.05	−0.50	0.866	9.3	9.6
1890	0.95	−0.50	−0.866	10.4	9.8
1900	1.00	0	0	12.3	11.7
1900	1.00	0	0	12.2	12.1
1880	1.00	−1.00	0	7.7	7.8
1910	1.05	0.50	0.866	6.2	5.8
1910	0.95	0.50	−0.866	11.3	11.6
1900	1.00	0	0	11.8	12.4
1900	1.00	0	0	12.7	12.0

The estimates $\hat{\beta}_1$,$\hat{\beta}_2$, and $\hat{\beta}_{12}$ are obtained directly:

$$\hat{\beta}_1 = \frac{-3.5000}{6} = -0.5833, \quad \hat{\beta}_2 = \frac{-10.5652}{6} = -1.7609, \quad \hat{\beta}_{12} = \frac{-4.1568}{1.5} = -2.7712 \tag{19.3.7}$$

Solving the remaining three equations, we obtain

$$\hat{\beta}_0 = 12.1500, \quad \hat{\beta}_{11} = -4.7500, \quad \hat{\beta}_{22} = -2.2833$$

Table 19.3.8 ANOVA table for second-order design.

Source	*DF*	*SS*	*MS*	*F*-ratio
First-order coefficient	2	20.6456	10.3228	
Second-order coefficient	3	80.0091	26.6697	208
Residual	14	1.7883	0.1277	
Lack of fit	1	0.6483	0.6483	7.40
Pure error	13	1.1400	0.0876	
Total	19	102.4480		

Hence, the fitted model is

$$\hat{y} = 12.15 - 0.58x_1 - 1.76x_2 - 4.75x_1^2 - 2.28x_2^2 - 2.77x_1x_2 \tag{19.3.8}$$

Some helpful computations for preparing the ANOVA table are

$$n = 20, \ \sum y = 200.8, \quad \sum y^2 = 2118.48, \quad \sum x_1 y = -3.5, \quad \sum x_2 y = -10.5652,$$

$$\sum x_1^2 y = 48.1, \quad \sum x_2^2 y = 55.5, \quad \sum x_1 x_2 y = -4.1568$$

Using some of the expressions in Table 19.5.1 and the above computational results, we obtain the ANOVA table shown in Table 19.3.8.

Note that from the above ANOVA table, we have that S^2, the pure error estimate of σ^2, is based on 13 degrees of freedom and is equal to

$$S^2 = 0.0876$$

The ratio of the lack of fit mean square to the pure error mean square is nearly twice the critical value of $F_{1,13;\ 0.05} = 4.67$, so that the fitted second-order model is declared adequate to represent the unknown function. This is because in practice, the lack of fit for F is two or three times the critical value before the model is declared inadequate to represent the data. Here, then, the decision is made to continue with the second-order polynomial model after recognizing that this fitted model may not be the best model proposed, but, may be useful for empirical approximation.

Using the residual sum of squares and its degrees of freedom, the new estimate of the variance is $S^{\ 2} = 1.7883/14 = 0.1277$ with 14 degrees of freedom (see Table 19.3.8). A test of the hypothesis that contributions of all second-order terms are zero is provided by using the test statistic $F = MS(\textit{second order})/MS(\textit{residual})$ which is observed to be $26.6697/0.1277 = 208$. Since $\text{Prob}\{F_{3,14} \geq 3.34\} = 0.05$, the hypothesis is rejected.

The fitted second-order model can now be used to determine the approximate contours of the response function. For example, the setting $\hat{y} = 10$ into (19.3.8) gives the contour in the (x_1, x_2) coordinate system

$$10 = 12.15 - 0.58x_1 - 1.76x_2 - 4.75x_1^2 - 2.28x_2^2 - 2.77x_1x_2$$

which is the equation of an ellipse. This contour and other contours for the estimated response are shown in Figure 19.3.10.

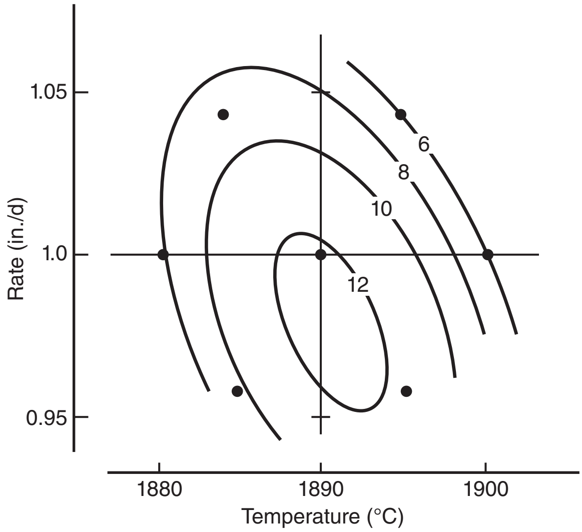


Figure 19.3.10 Contours of crystal purity as a function of temperature and rate of growth.

PRACTICE PROBLEM FOR SECTION 19.3

1. A chemical engineer wants to determine the optimal conditions of a chemical process. To achieve her goal, she wants to fit a second-order model to her data given below. The response Y is yield, and the design variables x_1 and x_2 are temperature and concentration (%), respectively. The whole experiment was carried out over a two months period. The first six experiments were performed during the first month, and the remaining six experiments were performed during the second month. Note that if in this experiment months are treated as blocks, then the blocks shown below are *orthogonal blocks*, i.e., block effects have no impact on the model coefficients.

x_1	x_2	Yield (Y)
−1	−1	36
1	−1	28
−1	1	26
1	1	38
0	0	33
0	0	35
−1.414	0	27
1.414	0	37
0	−1.414	31
0	1.414	34
0	0	29
0	0	32

(a) Fit a complete second-order model to these data.

(b) Construct an ANOVA table for these data. Estimate the error variance σ^2.

(c) Use the ANOVA table constructed in (b) to examine the significance of the second-order terms. Use $\alpha = 0.05$.

2. A pharmaceutical company wants to improve the effectiveness of plasma glucose-lowering drug used by Type-II diabetic patients by determining the proper mixture of the ingredients. The active and other ingredients used in this drug are hydrochloride, providone, and magnesium stearate. The company carried out the experiment given below. The response variable y indicates how much a tablet of 500 mg brings down the plasma glucose level (mg/DL) after the first week of use.

x_1	−1	1	−1	1	−1	1	−1	1	−1.682	1.682	0	0	0	0	0	0	0	0	0	0
x_2	−1	−1	1	1	−1	−1	1	1	0	0	−1.682	1.682	0	0	0	0	0	0	0	0
x_3	−1	−1	−1	−1	1	1	1	1	0	0	0	0	−1.682	1.682	0	0	0	0	0	0
y	33	26	50	30	29	47	24	37	37	45	48	35	42	48	41	45	33	37	27	47

(a) Fit a second-order model to these data.

(b) Construct an ANOVA table for these data. Estimate the error variance σ^2.

(c) Use the ANOVA table constructed in (b) to examine the significance of the second-order terms. Use $\alpha = .05$.

3. Refer to Problem 2. Consider a design that consists of the eight factorial points listed in Problem 2 and four center points. For the center points, use the last four data points on the list of Problem 2.

(a) Fit a first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the above data.

(b) Find the pure error and lack of fit mean square.

(c) Prepare an ANOVA table for these data and test the adequacy of the first-order model.

4. An industrial engineer used a hexagon design to examine a process development study. The hexagon design and some simulated data are given below.

1	0	16
−1	0	15
0.5	0.866	18
−0.5	0.866	16
0	0	22
−0.5	−0.866	25
0.5	−0.866	22
0	0	23
0	0	15
0	0	25
0	0	13
0	0	21

(a) Fit the complete second-order model $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$ to these data.
(b) Construct an ANOVA table for these data. Estimate the error variance σ^2.
(c) Use the ANOVA table constructed in (b) to examine the significance of the first-order and second-order terms. Use $\alpha = 0.05$.

5. Consider the design given below, which is nearly rotatable and has three orthogonal blocks. The following data were simulated by a design engineer.

Blocks	1						2					
X_1	−1	1	1	−1	0	0	1	−1	−1	1	0	0
X_2	−1	1	−1	1	0	0	−1	1	−1	1	0	0
X_3	−1	−1	1	1	0	0	−1	−1	1	1	0	0
Y	29	39	27	21	35	33	36	18	31	21	23	32

Blocks	3								
X_1	−1.732	1.732	0	0	0	0	0	0	0
X_2	0	0	−1.732	1.732	0	0	0	0	0
X_3	0	0	0	0	−1.732	1.732	0	0	0
Y	25	23	19	25	27	28	30	33	37

(a) Fit a complete second-order model to these data, that is $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$
(b) Construct an ANOVA table for these data. Estimate the error variance σ^2.
(c) Use the ANOVA table constructed in (b) to examine the significance of the first-order and second-order terms. Use $\alpha = 0.05$.

19.4 DETERMINATION OF OPTIMUM OR NEAR-OPTIMUM POINT

In practice, determination of levels of various factors always has some limitations, whether financial limitations, time limitations, or availability of the experimental material. This does have consequences that in practice the experimenter is dealing with a limited experimental region. A usual goal of experimentation is to find the minimum number of experimental points $(x_1, x_2, \ldots, x_k)$ within the experimental region at which the response Y is maximized (or minimized if Y measures, say, an unwanted byproduct). It is also important to explore the neighborhood of the optimum point. It sometimes happens one cannot always maintain the optimum point but may be able to maintain points on some neighborhood without having any significant effect on the response. Sometimes it can also happen that a slight shift away from the optimum point improves the overall product, while at the optimum point itself we may have optimality with respect only to a particular characteristic.

An exact method for finding the optimum point, especially when the experimental error is large, is to explore the whole region. But in practice, this is usually not possible. In the chemical industry, for instance, the problem of finding an optimum point is quite common, but it is almost impossible to explore the whole region since generating observations is usually quite expensive and time-consuming. Sometimes, in such investigations, we have some previous knowledge, such as the experimental error is usually small, and that experiments will be conducted sequentially. Further, due to small experimental error, small changes can be detected accurately and the experimenter can explore a small sub region adequately with only a few experiments. Since the experiments are planned sequentially, a technique can be developed that allows the use of the results obtained in one sub region to move to another sub region where the response can be larger or smaller, depending on whether we are seeking a maximum or minimum optimal point. This way, by successive applications of such a procedure, a stationary point, or at least a near-stationary point of optimum response, can usually be reached.

Friedman and Savage (1947) gave a sequential one-factor-at-a-time procedure when various factors are involved, usually known in the statistical literature as the *single-factor method*. It consists of first finding an optimum response by varying the levels of the first factor while keeping the levels of other factors at their initial values, say $x_2, x_3, \ldots, x_k$. Suppose now that the response was maximized when the first factor was at level x_1^0, say. Then, in the next experiment, the level of the first factor is fixed at x_1^0, the levels of the second factor are varied at various levels, and the remaining factors are kept fixed at their initial values. Now the optimum level of the second factor, say x_2^0, is found. This method is repeated with all other factors until we obtain an optimum point $(x_1^0, \ldots, x_k^0)$.

Box and Wilson (1951) proposed a method of locating the optimum and exploring the response surface in which many factors are varied at the same time. In their work, they proposed the use of the *path of steepest ascent* to get to a near-stationary region if the experimenter starts at a point far removed from it. When the experimenter comes near such a region, they described the use of certain composite designs that allows for the estimation of all the coefficients of the quadratic polynomial. We have already discussed these designs in our earlier part of this chapter. We now study methods to determine the optimum or near-optimum points.

19.4.1 The Method of Steepest Ascent

The method of steepest ascent is popular with experimenters and researchers. After it was proposed by Box and Wilson (1951), the method was further developed by Box, his collaborators, and others; some of these references are listed at the end of this book.

The steepest ascent method is a procedure in which the experimenter proceeds sequentially along the path of steepest ascent, that is, the path of maximum increase in response, or the path of steepest descent, that is, the path of maximum decrease in response. This method works more efficiently when the experimental error is small, and the lack of fit of a first-order model is not significant.

This method entails the starting of the experiment with a small design, for example, a 2^k or a fractional factorial design with some center points. At the end of this experiment two steps are taken: (i) a first-degree polynomial to approximate the response surface $f(x_1, x_2, \ldots, x_k)$ is fitted to the observed data; (ii) the linear approximation is tested to verify whether it fits the data adequately. Because we wish to test the adequacy of the first-order model, it is important to note that the experiment should be designed in such a way as to provide some degrees of freedom to measure the lack of fit and to estimate the experimental error or the pure error. The estimate of the experimental error is usually obtained either by replicating the runs or simply by adding some points at the center of the design. Further, if the fitted first-order model

$$\hat{y} = \hat{\beta}_0 x_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{19.4.1}$$

is adequate, then the center of the experimental region is moved to a new point that is chosen along a path toward where the maximum expected increase or decrease takes place. This path is called the *path of steepest ascent* or *steepest descent*, depending upon whether the desired response is a maximum or a minimum. This process is continued until there is no significant change in the response. At that point another experiment is performed to fit a first-order model. The lack-of-fit statistic for the first-order fit usually indicates that we have arrived in the vicinity of the optimum point.

The path along which we move is determined by

$$\frac{\partial f}{\partial x_i} = \hat{\beta}_i \quad i = 1, \ldots, k \tag{19.4.1a}$$

so that the change in x_i along the path of steepest ascent is proportional to $\hat{\beta}_i$. Note that if after the first set of experiments we find either that the first-order model is inadequate or that it is barely adequate with quite small regression coefficients, then we are in or near the stationary region. In this case, we do not proceed to calculate the path of steepest ascent, but rather add more points to the earlier design that will allow estimating the quadratic effects. We illustrate the use of method of determining the path of steepest ascent with the following example.

Example 19.4.1 (Experiment for maximizing the yield of a chemical) *A chemical engineer exposed a chemical A of certain percentage of concentration C to a certain temperature T for a certain time t observing the amount produced of another valuable chemical B. The experimenter was interested in finding the suitable concentration C, temperature T, and time t so that the yield of B is maximized. To achieve this aim, she performed an experiment using a design consisting of a* 2^3 *factorial and three points at the center, where the* 2^3 *factorial design has eight treatment combinations at the following factor levels,* $C =$ concentration *of A (%),* $t =$ times *(hours),* $T =$ temperature *(°C).*

C	t	T
25	4	250
30	6	300

The center points obtained by the various factor levels are generated at $C = 27.5$, $t = 5$, and $T = 275$.

Now, taking the center at the origin (0, 0, 0) of a new coordinate system, say x_1, x_2, x_3, and scaling the levels in the units of a factorial design, we find the eight combinations are $(\pm 1, \pm 1, \pm 1)$, with the relation between variables x_1, x_2, x_3, and the natural variables defined by

$$x_1 = \frac{C - 27.5}{2.5}, \quad x_2 = \frac{t - 5}{1}, \quad x_3 = \frac{T - 275}{25} \tag{19.4.2}$$

The data obtained by the first set of experiments are shown in Table 19.4.1. From this table, we have

$N = 11$, $\sum x_1^2 = \sum x_2^2 = \sum x_3^2 = 8$, $\sum x_1 y = 5.4$, $\sum x_2 y = 16.2$, $\sum x_3 y = 14.6$, $\sum y = 220$

Table 19.4.1 Data on a 2^3 factorial plus three center points.

x_1	x_2	x_3	Yield (g)
−1	−1	−1	15.3
1	−1	−1	17.5
−1	1	−1	18.9
1	1	−1	19.5
−1	−1	1	18.4
1	−1	1	19.2
−1	1	1	23.2
1	1	1	25.0
0	0	0	21.5
0	0	0	20.5
0	0	0	21.0

Hence, the estimates of the β_i's are

$$\hat{\beta}_0 = \frac{220}{11} = 20.0, \quad \hat{\beta}_1 = \frac{\sum x_1 y}{\sum x_1^2} = \frac{5.4}{8} = 0.675, \quad \hat{\beta}_2 = \frac{\sum x_2 y}{\sum x_2^2} = \frac{16.2}{8} = 2.025,$$

$$\hat{\beta}_3 = \frac{\sum x_3 y}{\sum x_3^2} = \frac{14.6}{8} = 1.825$$

Thus, the fitted first-order model is

$$\hat{Y} = 20 + 0.675x_1 + 2.025x_2 + 1.825x_3 \tag{19.4.3}$$

Further, using the techniques discussed earlier in this chapter, we can easily verify that the ANOVA table fitting the first-order model based on the data in Table 19.4.1 is as shown in Table 19.4.2.

Table 19.4.2 ANOVA table for the data in Table 19.4.1.

Source	*SS*	*DF*	*MS*	*F*-ratio
β_1	3.645	1		
β_2	32.805	1		
β_3	26.645	1		
Lack of fit	8.145	5	1.629	6.516
Pure error	0.500	2	0.25	
Total	71.740	10		

The F-ratio for testing "no-lack of fit" is 6.516, which is less than the upper 5% point of the $F_{5,2}$ distribution $((F_{5,2;\ 0.05} \geq 19.296) = 0.05)$, so that the fit of the first-degree polynomial is deemed quite adequate. Also, the regression coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are not very small. Thus, our next step is to determine the path of steepest ascent. From the fitted equation, we have that the changes in the x_i's, $i = 1$, 2, and 3 (in the units of the design) along the *path of steepest ascent* are proportional to

$$\hat{\beta}_1 = 0.675, \quad \hat{\beta}_2 = 2.025, \quad \hat{\beta}_3 = 1.825$$

respectively. Thus, in the *original units*, the changes along the path in the original variables (see (19.4.2)) are proportional to

$$C = 0.675 \times 2.5 = 1.6875; \; t = 2.025 \times 1 = 2.025; \; T = 1.825 \times 25 = 45.625.$$

Now we select one of the variables, say concentration, as the standard variable and then calculate the *changes in the other variables that would correspond to a 1% change in concentration.* Thus, the changes in C, t and T are 1, $1.2 = 2.025/1.6875$, and $27.037 = 45.625/1.6875$, respectively.

Now, to obtain the path of steepest ascent, we start with the center point, which was at the origin. From above, we have that for each 1% increase of the concentration, the time is increased by 1.2 hours and the temperature is increased by 27.037 °C. Table 19.4.3 gives the various points on the path of steepest ascent. The predicted values of $\hat{y}$ at these points as calculated from the fitted first-order model and are given in Table 19.4.4. Note that to calculate the predicted values, all the levels must be first converted into design units.

Table 19.4.3 Points on the steepest ascent.

	C	t	T
Initial point (center point)	27.5	5	275
1	28.5	6.2	302.037
2	29.5	7.4	329.074
3	30.5	8.6	356.111
4	31.5	9.8	383.148

The predicted responses $\hat{y}$ at these points are given in Table 19.4.4.

Table 19.4.4 Predicted responses at the points along the path of steepest ascent (coded units).

	x_1	x_2	x_3	$\hat{Y}$
Initial point	0	0	0	20
1	0.40	1.20	1.0815	24.6740
2	0.80	2.40	2.1630	29.3480
3	1.20	3.60	3.2445	34.0220
4	1.60	4.80	4.3260	38.6960

Then, the next step is to conduct a single experiment at the point expressed in experimental units of (x_1, x_2, x_3) as (0.40, 1.2, 1.0815). If the actual observed response for this point when generated is close to the predicted response, a further ascent is made along the same path. This procedure is continued until the actual yield differs substantially from the predicted yield; that is, we continue until there is a significant increase in the response. At this point, a new 2^k factorial experiment is conducted with its center at the last point of the path. If a first-order model can be fitted adequately, then again, we determine the direction of the new path of steepest ascent and repeat the above process (again, note that if the fitted model is barely adequate and the regression coefficients of the new fitted model are quite small, we usually do not need to determine the new path of steepest ascent; rather, we explore that region by conducting additional experiments). Finally, a situation is reached in which the 2^k factorial gives one of the following situations:

1. The first-order model seems to fit adequately, but all the regression coefficients (the $\hat{\beta}_i$'s) are very small. This means a plateau is reached. At this point, the curvature of the surface should be considered by conducting additional experiments.
2. The lack of fit test shows that the first-order model is not adequate. This implies that the curvature of the surface should be considered and additional experiments should be conducted.
3. The first-order model seems to fit adequately and all the regression coefficients $\hat{\beta}_i$'s are not very small; then a new steepest ascent path should be determined and the whole process should be repeated.

To explore the curvature, a second-order CCD is performed. The response surface is then explored by transferring the response surface to its *canonical* form. We study the canonical form next in Section 19.4.2.

19.4.2 Analysis of a Fitted Second-Order Response Surface

Earlier in this chapter, we saw that if the fitted response function is a second-degree polynomial in two variables, then the fitted response surface may be represented by a mound, basin, saddle, stationary ridge, rising ridge, or falling ridge (the contours of these surfaces are plotted in Figure 19.1.3). In this section, we study how to determine the nature of a stationary point and the nature of the response surface. This determination is done by reducing the fitted second-degree polynomial to its standard form, called the *canonical* form.

The fitted second-degree polynomial is reduced to its canonical form by first shifting the origin to the stationary point (x^0) and then using an orthogonal transformation. This, of course, is possible for a polynomial in any number of variables, and we illustrate the procedure by considering a two-variable case.

The general form of a second-degree polynomial in two variables "representing a fitted surface" is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{22} x_2^2 + \hat{\beta}_{12} x_1 x_2 \tag{19.4.4}$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{12}$ are the least-squares estimators of the corresponding coefficients in the second-degree polynomial model that is fitted, which is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{19.4.5}$$

If we shift the origin of the x-coordinates to the stationary point (x_1^0, x_2^0), that is, we let

$$x_1 = u_1 + x_1^0, \quad x_2 = u_2 + x_2^0$$

Here u_1, u_2 are the new variables, and the polynomial (19.4.4) reduces to

$$\hat{Y} = \hat{y}_0 + \hat{\beta}_{11} u_1^2 + \hat{\beta}_{22} u_2^2 + \hat{\beta}_{12} u_1 u_2 \tag{19.4.6}$$

since

$$\left(\frac{\partial \hat{Y}}{\partial x_1}\right)_{x=x^0} = \left(\frac{\partial \hat{Y}}{\partial x_2}\right)_{x=x^0} = 0$$

and $\hat{y}_0$ is the value of the estimated response at the stationary point (x_1^0, x_2^0). In matrix notation (19.4.6) may be written as

$$\hat{Y} = \hat{y}_0 + U'BU \tag{19.4.7}$$

where

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad B = \begin{bmatrix} \hat{\beta}_{11} & \frac{1}{2}\hat{\beta}_{12} \\ \frac{1}{2}\hat{\beta}_{21} & \hat{\beta}_{22} \end{bmatrix}$$

We note that B defined above is a real symmetric matrix. As is well known, there exists an orthogonal matrix, say P, that diagonalizes B, that is

$$PBP' = D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

where $(\lambda_1,\ \lambda_2)$ are the eigenvalues of the matrix B. Hence, since P is orthogonal we can easily see that

$$B = P'DP$$

We then have that

$$U'BU = U'(P'DP)U = (PU)'D(PU) \tag{19.4.8}$$

Now if we transform U orthogonally, say from U to V, where $V = PU$, we then have that (19.4.7) may be written as (see (19.4.8))

$$\hat{Y} - \hat{y}_0 = V'DV = (v_1, v_2)' \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda_1 v_1^2 + \lambda_2 v_2^2 \tag{19.4.9}$$

This is the canonical form of the polynomial (19.4.4).

The expression in (19.4.9) may now be interpreted as follows: clearly, the right side of (19.4.9) expresses the change in the estimated response as we move away from the stationary point $(v_1, v_2) = (0{,}0)$ with value $\hat{Y} = \hat{y}_0$ to some other point (v_1, v_2) with value $\hat{y}(v_1, v_2)$. Moreover, (19.4.9) contains only the squared terms in the variables v_1 and v_2, and any change in $\hat{y}$ is completely determined by the λ's. Thus, we have the following:

(i) If both λ_1 and λ_2 are negative, then any move away from the stationary point results in a decrease of $\hat{Y}$, which implies that the stationary point (x_1^0, x_2^0) is a point of maximum response and the response surface is a mound.

(ii) If both λ_1 and λ_2 are positive, then any move away from the stationary point results in an increase of $\hat{Y}$, which implies that the stationary point (x_1^0, x_2^0) is a point of minimum response and the response surface is a basin.

(iii) If one λ is positive and the other is negative, then the stationary point is a saddle point and the response surface is a saddle. With the saddle point, the stationary point is neither a minimum nor a maximum point, and we refer to it as a minmax point. We may get an increase or decrease in $\hat{Y}$ as we move away from the stationary point, depending upon the direction of our move. Further, if one of the λ's is either zero or nearly zero, then we have a surface that is a stationary ridge or a near-stationary ridge, in which case we do not have a single maximum or minimum response. For example, if the stationary point is a maximum, but one of the λ's is very small, then the stationary point is not the unique maximum point and, in fact, all points along a certain straight-line are maximums. Also, if the stationary point is not in the vicinity of the experimental region, but it is far removed from it, then we have a situation of rising ridge or falling ridge, depending upon whether the response increases or decreases as we move away from the stationary point.

The case for more than two variables can be studied in the same way. Thus, for example,

(i) if all the λ's are negative, then the stationary point is a maximum point.
(ii) if all the λ's are positive, then the stationary point is a minimum point.
(iii) if λ's differ in sign, then the stationary point is a saddle point.
(iv) if one or more λ's are either zero or very small, then we have a stationary ridge.

As a further illustration, we consider the following example with three variables (x_1, x_2, x_3).

Example 19.4.2 (Determining the type of response surface) *Consider the data obtained by using a rotatable CCD with three factors A, B, and C. The design plan and the simulated data obtained are presented in Table 19.4.5. The experiment was carried out in random order.*

Solution: Again, note that the above design is rotatable, but not orthogonal (see Definition 19.3.1). Thus, using the technique discussed in Section 19.5, we find ($N = 20$)

$$\hat{\beta}_0 = 23.85, \quad \hat{\beta}_1 = 0.78, \quad \hat{\beta}_2 = 0.03, \quad \hat{\beta}_3 = 0.76, \quad \hat{\beta}_{11} = -1.72,$$
$$\hat{\beta}_{22} = -2.53, \quad \hat{\beta}_{33} = -0.80, \quad \hat{\beta}_{12} = -1.63, \quad \hat{\beta}_{13} = 1.00, \quad \hat{\beta}_{23} = 0.25$$

Thus, the second-order fitted model is

$$\hat{Y} = 23.85 + 0.78x_1 + 0.03x_2 + 0.76x_3 - 1.72x_1^2 - 2.53x_2^2 - 0.80x_3^2 - 1.63x_1x_2 + x_1x_3 + 0.25x_2x_3$$

The ANOVA table for the data in Table 19.4.5 is shown in Table 19.4.6.

From the above analysis, we see that the fit of the second-order model is quite adequate. We now determine the nature of the stationary point and the fitted response surface.

Table 19.4.5 Design plan and the data obtained using this plan.

x_1	x_2	x_3	Y
−1	−1	−1	18.5
1	−1	−1	23.0
−1	1	−1	21.0
1	1	−1	16.5
−1	−1	1	19.0
1	−1	1	25.0
−1	1	1	20.0
1	1	1	22.0
0	0	0	23.4
0	0	0	26.2
0	0	0	24.5
0	0	0	20.6
0	0	0	25.3
0	0	0	24.0
−1.682	0	0	15.6
1.682	0	0	17.2
0	−1.682	0	12.2
0	1.682	0	16.0
0	0	−1.682	18.0
0	0	1.682	20.0

Table 19.4.6 ANOVA table for the data in Table 19.4.5.

Source	*DF*	*SS*	*MS*	*F*-ratio
First-order terms	3	16.24	5.41	1.45
Second-order terms	6	156.14	26.02	6.96
Lack of fit	5	82.96	16.59	4.44
Pure error	5	18.70	3.74	
Total	19	274.04		

We first find the stationary point (x_1^0, x_2^0, x_3^0). It is well known that the stationary point is the solution of the set of equations

$$\frac{\partial \hat{Y}}{\partial x_1} = 0, \quad \frac{\partial \hat{Y}}{\partial x_2} = 0, \quad \frac{\partial \hat{Y}}{\partial x_3} = 0$$

The three equations are

$$0.78 - 3.44x_1^0 - 1.63x_2^0 + x_3^0 = 0$$

$$0.03 - 5.06x_2^0 - 1.63x_1^0 + 0.25x_3^0 = 0$$

$$0.76 - 1.60x_3^0 + x_1^0 + 0.25x_2^0 = 0$$

In matrix notation, this set of equations can be written as

$$\begin{bmatrix} -3.44 & -1.63 & 1.00 \\ -1.63 & -5.06 & 0.25 \\ 1.00 & 0.25 & -1.60 \end{bmatrix} \begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{bmatrix} = \begin{bmatrix} -0.78 \\ -0.03 \\ -0.76 \end{bmatrix}$$

or

$$\begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{bmatrix} = \begin{bmatrix} -3.44 & -1.63 & 1.00 \\ -1.63 & -5.06 & 0.25 \\ 1.00 & 0.25 & -1.60 \end{bmatrix}^{-1} \begin{bmatrix} -0.78 \\ -0.03 \\ -0.76 \end{bmatrix}$$

or

$$\begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{bmatrix} = \begin{bmatrix} -0.419741 & 0.123203 & -0.243087 \\ 0.123203 & -0.235329 & 0.040232 \\ -0.243087 & 0.040232 & -0.770643 \end{bmatrix} \times \begin{bmatrix} -0.78 \\ -0.03 \\ -0.76 \end{bmatrix} = \begin{bmatrix} 0.508448 \\ -0.119615 \\ 0.774090 \end{bmatrix}$$

Thus, the stationary point is (0.51, −0.12, 0.77).

The value of the estimated response $\hat{Y}$ at the stationary point is

$$\hat{y}_0 = 24.341$$

Thus, if we now shift our origins to the stationary point, that is

$$x_1 = u_1 + 0.51, \quad x_2 = u_2 - 0.12, \quad x_3 = u_3 + 0.77$$

where (u_1, u_2, u_3) are the new coordinates. The fitted polynomial can then be written as

$$\hat{Y} = 24.341 - 1.72u_1^2 - 2.53u_2^2 - 0.80u_3^2 - 1.63u_1u_2 + u_1u_3 + 0.25u_2u_3$$

This in turn can be reduced to its canonical form by determining the eigenvalues of the matrix

$$B = \begin{bmatrix} -1.720 & -0.815 & 0.500 \\ -0.815 & -2.530 & 0.125 \\ 0.500 & 0.125 & -0.800 \end{bmatrix}$$

The eigenvalues of this matrix are found to be $-3.09, -1.40, -0.55$.

Thus, the canonical form of the second-order fitted model or the response function is

$$\hat{Y} - 24.341 = -3.09u_1^2 - 1.40u_2^2 - 0.55u_3^2$$

Since all the eigenvalues are negative, the stationary point is a maximum point and the response surface is a mound.

MINITAB

Note that we can find the eigenvalues of the matrix B using **MINITAB** as follows:

1. Enter the matrix data in the worksheet as columns.
2. Select **Data > Copy > Columns to Matrix.**

3. In the dialog box, enter the columns in the box under **Copy from columns**, and the name of the matrix, say, M_1 in the box under **in current worksheet, in matrix** and click **OK**.
4. Select **Calc** > **Matrices** > **Eigen Analysis.** In the dialog box **Eigen Analysis** enter the name of the matrix in the box next to **Analyze matrix:**, enter the column where you would like to store the eigen values in the box next to **Column of eigenvalues:** and click **OK**. The eigenvalues will appear in the desired column of the worksheet.

USING R

Solution The 'eigen()' function in R can be used to obtain both eigenvalues and eigenvectors of the matrix M1 as follows.

```
M1 = matrix(c(-1.720, -0.815, 0.500, -0.815, -2.530, 0.125, 0.500, 0.125, -0.800),
3, 3, byrow = 'TRUE')
eigen(M1)
```

PRACTICE PROBLEMS FOR SECTION 19.4

1. Refer to Problem 1 of Section 19.3. Determine the canonical form of the fitted second-order model. Find whether the stationary point is a maximum, a minimum, a saddle point, or a stationary ridge.
2. Refer to Problem 3 Section 19.3. Find five points on the path of steepest ascent in terms of the coded variables.
3. Refer to Problem 2 of Section 19.3. Determine the canonical form of the fitted second-order model. Find whether the stationary point is a maximum, a minimum, a saddle point, or a stationary ridge.
4. Refer to Problem 5 of Section 19.3. Determine the canonical form of the fitted second-order model. Find whether the stationary point is a maximum, a minimum, a saddle point, or a stationary ridge.
5. Refer to Problem 5 of Section 19.3. Construct a design to fit a first-order model using the factorial points and center points given in blocks 1 and 2 of Problem 5 of Section 19.3. Then use the corresponding data points to do the following:
   (a) Fit a first-order model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ to the above data.
   (b) Find four points on the steepest ascent in terms of the coded variables.

19.5 ANOVA TABLE FOR A SECOND-ORDER MODEL

In order to fit the second-order model in k variables, we first note that there are $1 + k + k + k(k-1)/2 = [(k+1) \times (k+2)]/2$ unknown coefficients in the model

$$\eta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum \sum_{i<j}^{k} \beta_{ij} x_i x_j \tag{19.5.1}$$

for a total of $\binom{k+2}{2}$ coefficients to be estimated. Further, in order to estimate quadratic coefficients, a minimum of three levels of each of the variables ξ_i must be used. This would seem to suggest that the 3^k factorial designs would be useful for securing data for second-order model fitting. However, many other appropriate designs exist, many of them with valuable properties we have discussed in this chapter, requiring fewer experimental points.

For $k = 2$ the second-order model is $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$. The normal equations associated with this model are

$$\begin{aligned}
N\hat{\beta}_0 + [1]\hat{\beta}_1 + [2]\hat{\beta}_2 + [11]\hat{\beta}_{11} + [22]\hat{\beta}_{22} + [12]\hat{\beta}_{12} &= \sum y_i \\
[1]\hat{\beta}_0 + [11]\hat{\beta}_1 + [12]\hat{\beta}_2 + [111]\hat{\beta}_{11} + [122]\hat{\beta}_{22} + [112]\hat{\beta}_{12} &= \sum x_{i1} y_i \\
[2]\hat{\beta}_0 + [12]\hat{\beta}_1 + [22]\hat{\beta}_2 + [112]\hat{\beta}_{11} + [222]\hat{\beta}_{22} + [122]\hat{\beta}_{12} &= \sum x_{i2} y_i \\
[11]\hat{\beta}_0 + [111]\hat{\beta}_1 + [112]\hat{\beta}_2 + [1111]\hat{\beta}_{11} + [1122]\hat{\beta}_{22} + [1112]\hat{\beta}_{12} &= \sum x_{i1}^2 y_i \\
[22]\hat{\beta}_0 + [122]\hat{\beta}_1 + [222]\hat{\beta}_2 + [1122]\hat{\beta}_{11} + [2222]\hat{\beta}_{22} + [1222]\hat{\beta}_{12} &= \sum x_{i2}^2 y_i \\
[12]\hat{\beta}_0 + [112]\hat{\beta}_1 + [122]\hat{\beta}_2 + [1112]\hat{\beta}_{11} + [1222]\hat{\beta}_{22} + [1122]\hat{\beta}_{12} &= \sum x_{i1} x_{i2} y_i
\end{aligned} \tag{19.5.2}$$

where $\sum_{i=1}^n x_{iu} = [u]$, $\sum_{i=1}^n x_{iu}^2 = [uu]$, $\sum_{i=1}^n x_{iu} x_{iw} = [uw]$, $\sum_{i=1}^n x_{iu}^2 x_{iw} = [uuw]$, $\sum_{i=1}^n x_{iu}^2 x_{iw}^2 = [uuww]$, $\sum_{i=1}^n x_{iu}^4 = [uuuu]$, etc. Here n denotes the common number of entries used for factors x_u and x_w.

The generalization of these normal equations for $k > 2$ should be obvious. A characteristic of the symmetrical designs considered in this chapter is that the mixed second sums $[uw]$, all the odd sums $[u], [uuu], [uuw]$ and all fourth-order sums of the form $[uuuw]$ are zero, leading to a simplification of the normal equations. For these designs, we have directly $\hat{\beta}_u = \sum x_{iu} y_i / [uu]$, and $\hat{\beta}_{uw} = \sum x_{iu} x_{iw} y_i / [uuww]$, $u \neq w \neq 0$. The k second-order coefficients $\hat{\beta}_{uu}$, along with $\hat{\beta}_0$ are readily obtained by ordinary algebra.

The ANOVA for this case appears in Table 19.5.1.

From Table 19.5.1, we have that $S_e^2 = SSE/\nu$. Of course, if “no lack of fit” is not rejected, then we use $MSE = SS(Resid)/\left[N - \binom{k+2}{2}\right]$ to estimate σ^2, etc. Whenever replicated runs can be used to provide a pure error sum of squares with v degrees of freedom, the residual sum of squares SS may be partitioned into two parts:

$$\text{Residual } SS = \text{pure error } SS + \text{lack-of-fit } SS$$

The lack-of-fit and error mean squares provides a measure of the adequacy of the fitted model, which for $k = 2$, is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{22} x_2^2 + \hat{\beta}_{12} x_1 x_2 \tag{19.5.3}$$

If the model is accepted as adequate to represent the unknown response function, (19.5.3) can be used to provide an approximate map of the response over the experimental region. The contours of the fitted surface are obtained by setting $\hat{Y}$ equal to selected specific values of the response and plotting the resulting second-degree equations in the coordinate system of x_1 and x_2. The contours will be concentric ellipses, or hyperbolas, or

Table 19.5.1 Analysis of variance table for a second-order model in k variables $\eta = \beta_0 + \sum_{u=1}^{k} \beta_u x_u + \sum_{u=1}^{k} \beta_{uu} x_u^2 + \sum_u \sum_{w:1 \leq u < w \leq k} \beta_{uw} x_u x_w$.

Source	*SS*	Formulas	*DF*
Total corrected	SST	$\sum y^2 - (\sum y)^2/N = SS_{total}$	$N-1$
First-order coefficients	$SS(\hat{\beta}_u)$	$\sum_{u=1}^{k} \hat{\beta}_u (\sum_{i=1}^{n} x_{iu} y_i)$	k
Second-order coefficients	$SS(\hat{\beta}_{uu})$	$\sum_{u=1}^{k} \hat{\beta}_{uu} (\sum_{i=1}^{n} x_{iu}^2 y_i)$	k
Two-factor interaction coefficients	$SS(\hat{\beta}_{uw})$	$\sum \sum_{1 \leq u < w \leq k}^{k} \hat{\beta}_{uw} \left(\sum_{i=1}^{n} x_{iu} x_{iw} y_i\right)$	$k(k-1)/2$
Residual *SS*	SS(Res.)	$SS_{total} - SS(\hat{\beta}_u) - SS(\hat{\beta}_{uu}) - SS(\hat{\beta}_{uw})$	$N - \binom{k+2}{2}$
Lack of fit *SS*	SS_L	SS(Res.) $- SSE$	$N - \binom{k+2}{2} - \nu$
Pure error *SS*	SS_E	SSE	ν

even straight-lines, as illustrated in Figure 19.1.3. For further investigation of *RSM* the reader is referred to Box and Draper (1987), Montgomery (2009), Myers and Montgomery (2002), and Khuri and Cornell (1996).

19.6 CASE STUDIES

Case Study 1 Using the data in Chapter 18 of Case Study 2 of Section 18.6, do the following:

(a) Fit a first-order model $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.
(b) Test the adequacy of the fitted model in (a) using $\alpha = 0.05$.
(c) Find five points on the path of steepest ascent.

Case Study 2: (Rayon Whiteness in Fabric Study)[1]
Rayon whiteness is an important factor for scientists dealing with fabric. Whiteness is affected by pulp quality and other factors such as temperature, °C (x_1); cascade acid concentration, % (x_2); water temperature, °C (x_3); sulfide concentration, % (x_4); amount of chlorine bleach, lb/min (x_5). An experiment using these factors is planned to study this problem. To perform this experiment, a rotatable CCD consisting of a one-half fraction of

[1] *Source*: Myers and Montgomery (1995), used with permission.

a 2^5 design, 10 axial points and five center points is used. The responses “whiteness” and all the treatments are given in Table 19.6.1.

Table 19.6.1 Treatment matrix with responses.

2^{5-1} replication						Axial and center points					
x_1	x_2	x_3	x_4	x_5	Response	x_1	x_2	x_3	x_4	x_5	Response
−1	−1	−1	−1	−1	71.5	−2	0	0	0	0	80.2
1	1	−1	−1	−1	76.0	2	0	0	0	0	84.1
1	−1	1	−1	−1	79.9	0	−2	0	0	0	77.2
1	−1	−1	1	−1	83.5	0	2	0	0	0	85.1
1	−1	−1	−1	1	89.5	0	0	−2	0	0	71.5
−1	1	1	−1	−1	84.2	0	0	2	0	0	84.4
−1	1	−1	1	−1	85.3	0	0	0	−2	0	77.5
−1	1	−1	−1	1	94.5	0	0	0	2	0	79.2
−1	−1	1	1	−1	89.4	0	0	0	0	−2	71.0
−1	−1	1	−1	1	97.5	0	0	0	0	2	90.2
−1	−1	−1	1	1	103.2	0	0	0	0	0	72.1
1	1	1	1	−1	108.7	0	0	0	0	0	72.0
1	1	1	−1	1	115.2	0	0	0	0	0	72.4
1	1	−1	1	1	111.5	0	0	0	0	0	71.7
1	−1	1	1	1	102.3	0	0	0	0	0	72.8
−1	1	1	1	1	108.1						

The coding of the design variables is as given below. $x_1 = \dfrac{\text{temp} - 45}{10}$; $x_2 = \dfrac{\text{casdconc} - 0.5}{0.2}$; $x_3 = \dfrac{\text{watertemp} - 85}{3}$; $x_4 = \dfrac{\text{sulfconc} - 0.25}{0.05}$; $x_5 = \dfrac{\text{chbl} - 0.4}{0.1}$

(a) Fit a second-order model to the response data in Table 19.6.1.

(b) Determine the canonical form of the model fitted in (a). Give the type of the response surface.

(c) What operating conditions should be used if it is important to minimize the whiteness?

19.7 USING JMP

This section is not included here but is available for download from the book website: www.wiley.com/college/gupta/statistics2e.

Review Practice Problems

1. In Problem 1 of Section 19.2, let levels in actual units for the three factors be as given below

X_1	X_2	X_3
122	24	6
132	28	8
142	32	10

Find five points on the path of steepest ascent and then determine the corresponding predicted responses. *Note*: Standardized variables $(x_1,\ x_2,\ x_3)$ are obtained from $x_1 = (X_1 - 132)/10,\ x_2 = (X_2 - 28)/4,\ x_3 = (X_3 - 8)/2$.

2. Penicillin production requires a fermentation step that must be done in batches. A difficulty in the production of successive batches is the nutrient, corn steep liquor, which varies. A study was begun to determine whether changes in temperature and pH might increase the penicillin yields for a new set of fermenters. A 2^2 factorial design was employed, and a new batch of corn steep liquor employed for each set of four runs. The results are displayed below.

Design		Penicillin yields				
(pH)	(Temperature)	Corn steep liquor batches				
x_1	x_2	1	2	3	4	5
−1	−1	40	35	28	27	33
1	−1	95	80	94	76	83
−1	1	66	50	48	45	61
1	1	124	98	105	96	100

(a) Fit a first-order model to the response function.
(b) Assuming the postulated first-order model is appropriate, plot the contours of the best fitting plane.
(c) Plot the path of steepest ascent.

3. In a study of the breaking strength of concrete cylinders, it was decided to map breaking strength as a function of x_1: hours in mold, and x_2: age at test. The octagon

design was used (with all these design points on a circle) plus four runs at (0, 0). The runs of this design produced the observations shown in the table below.

Run number	x_1	x_2	y
1	$-\sqrt{2}$	0	77
2	0	0	95
3	0	0	94
4	−1	−1	84
5	1	1	95
6	0	0	93
7	0	$-\sqrt{2}$	76
8	−1	1	79
9	0	0	96
10	0	$\sqrt{2}$	88
11	1	−1	78
12	$\sqrt{2}$	0	84

(a) Plot the design points and at each point record the corresponding observations.
(b) Determine whether a first-order model appears adequate to represent the response function. Use $\alpha = 0.05$.
(c) Now fit a second-order model to the data.
(d) Determine the nature of the stationary point and the nature of the fitted response function.

4. Suppose an experiment using a hexagonal design with center points yields the observations as shown below.
(a) Plot the design coordinates and at each point records the corresponding observations.
(b) Fit a first-order model.
(c) Does the model adequately represent the data? Use $\alpha = 0.05$.
(d) Are both variables x_1 and x_2 necessary to explain the data?
(e) Sketch the fitted response function.

Run number	x_1	x_2	y
1	0	0	4.2
2	−0.500	0.866	3.7
3	0	0	4.4
4	0.500	−0.866	4.8
5	1	0	7.4
6	0	0	3.9
7	−0.500	−0.866	1.8
8	−1	0	1.1
9	0.500	0.866	6.8
10	0	0	4.2

5. The experimental design and data are shown below,
   (a) Plot the design and at each point record the corresponding observations.
   (b) Consider only the first eight runs and analyze this portion of the data as a replicated 2^2 factorial design. Does a first-order model adequately represent these data? Use $\alpha = 0.05$.
   (c) Using all the data, fit a second-order model.

Run number	x_1	x_2	y
1	-1	-1	45.9
2	-1	-1	53.3
3	-1	1	57.5
4	-1	1	58.8
5	1	-1	60.6
6	1	-1	58.0
7	1	1	58.6
8	1	1	52.6
9	0	0	56.9
10	2	0	55.4
11	-2	0	46.9
12	0	2	57.5
13	0	-2	55.0
14	0	0	58.9
15	0	0	50.3

6. A 3^2 factorial design was employed in a study to determine the best freezing conditions for orange juice, the response measured being the percent natural vitamin B remaining after eight weeks. The two variables are x_1: depth of freeze in °C, and x_2: rate of freeze. The recorded responses are given below.

x_1	x_2	y
0	0	81.5
1	1	75.8
0	1	80.2
0	-1	79.2
-1	1	76.0
1	1	74.3
0	0	81.3
1	-1	80.1
1	0	79.1
-1	-1	70.2
-1	0	75.2
-1	-1	71.7
-1	1	76.2
1	-1	81.0

(a) Plot the design and at each point, record the corresponding observations.
(b) Fit a second-order model.
(c) Determine whether the model adequately represents the data. State α.

7. Consider the design T, the principal block in a one-half fraction of a 2^3 factorial structure defined by the sentence $I = -ABC$.

$$T = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$$

Suppose now we fit the model with $x_0 = 1$, namely

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

but the true model is

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

(a) Show that $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ are biased estimators of $\beta_1, \beta_2,$ and β_3, respectively.
(b) In part (a), find the bias for each estimator.

8. In Problem 7, consider now the complete replication of a 2^3 factorial design to fit the model

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 \\ + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

(a) Show that $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{23}$, are unbiased estimators of $\beta_1, \beta_2, \ldots, \beta_{23}$, respectively.
(b) Show that $\hat{\beta}_0$ is a biased estimator of β_0. Find the bias in $\hat{\beta}_0$.
(c) Explain how you can test the curvature by just adding some center points to the complete 2^3 factorial design. (Curvature is measured by $\gamma = \beta_{11} + \beta_{22} + \beta_{33}$, so that we wish to test $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$.)

9. An experiment is performed to study the effects of three variables (A, B, and C) on the life of a cutting tool by employing a CCD. The variables are A, cutting tool speed (ft/min); B, the feed rate (in./rev); and C, the depth of cut (inches). (This study is fully reported in Wu (1964).) The settings of the three controlled variables A, B, and C, the equivalent settings of the design variables x_1, x_2, and x_3, and the observed response y (ln(tool life measured in minutes)) are listed in the table below. It was decided to fit a second-order model to the response function. The design employed was a CCD consisting of a 2^3 factorial with the center point replicated four times, plus a twice-replicated star design.

A	B	C	Design			Tool life
Speed: fpm	Feed: ipm	Depth: in.	x_1	x_2	x_3	$y = \ln(\text{min})$
330	0.01	0.04	−1	−1	−1	5.08
700	0.01	0.04	1	−1	−1	3.61
330	0.02	0.04	−1	1	−1	5.11
700	0.02	0.04	1	1	−1	3.30
330	0.01	0.10	−1	−1	1	5.15
700	0.01	0.10	1	−1	1	3.56
330	0.02	0.10	−1	1	1	4.79
700	0.02	0.10	1	1	1	2.89
515	0.015	0.07	0	0	0	4.19
515	0.015	0.07	0	0	0	4.42
515	0.015	0.07	0	0	0	4.26
515	0.015	0.07	0	0	0	4.41
145	0.015	0.07	−2	0	0	5.48
885	0.015	0.07	2	0	0	2.64
515	0.005	0.07	0	−2	0	4.70
515	0.025	0.07	0	2	0	3.99
515	0.015	0.01	0	0	−2	4.60
515	0.015	0.13	0	0	2	4.25
145	0.015	0.07	−2	0	0	5.41
885	0.015	0.07	2	0	0	2.71
515	0.005	0.07	0	−2	0	4.53
515	0.025	0.07	0	2	0	3.74
515	0.015	0.01	0	0	−2	4.66
515	0.015	0.13	0	0	2	4.17

(a) Use the first 12 data points given above to fit a first-order model.
(b) Estimate the error variance σ^2 with three degrees of freedom. Explain which data points provide these degrees of freedom.

10. Refer to Problem 9 and do the following:
(a) Use the entire data set of Problem 9 to fit a second-order model.
(b) Estimate the error variance σ^2 with nine degrees of freedom. Explain which data points contribute to these nine degrees of freedom.
(c) Use the estimate of the error variance σ^2 with nine degrees freedom in (b) to test the adequacy of the model you fitted in (a).

11. Divide the Residual sum of squares with 13 (= 23 − 10) degrees of freedom in Problem 10 above into two parts: (i) lack of fit with four degrees of freedom (ii) pure error sum of squares with nine degrees of freedom. Use these sums of squares to test the lack of fit.

12. For the fitted second-order model in Problem 10, find the nature of the stationary point and of the fitted response surface.

13. Consider a first-order design consisting of (i) $2^4 = 16$ factorial points of a complete replication of a 2^4 factorial design and (ii) n_c center points. Let $\bar{Y}$, $\bar{Y}_c$ be the average responses at the factorial points and center points, respectively. Suppose the true response surface is a quadratic polynomial that includes the x_i^2 terms and the $x_i x_j$ terms. Under this assumption show that the difference of the two averages $(\bar{Y} - \bar{Y}_c)$ estimates the sum of the pure quadratic coefficients $\beta_{11} + \beta_{22} + \beta_{33} + \beta_{44}$. Further, show that the sum of squares due to these coefficients with one degree of freedom is given by $\frac{n\,n_c}{n + n_c}(\bar{Y} - \bar{Y}_c)^2$, here $n = 2^4 = 16$. How do you use this sum of squares to test a hypothesis that the response surface has some curvature?

14. Suppose that a fitted second-order response surface model is the following:

$$\hat{Y} = 15.4 + 0.5x_1 - 1.2x_2 + 0.85x_3 + 2.6x_1x_2 - 1.8x_1x_3 + 2.1x_2x_3 + 3.2x_1^2 + 1.4x_2^2 + 2.7x_3^2$$

Reduce this fitted model to its canonical form and then describe the nature of the fitted response surface.

15. Consider the following first-order fitted model.

$$\hat{Y} = 50 + 2.2x_1 - 1.2x_2 + 2.1x_3$$

Find the path of steepest ascent. The variables are coded using the standard design units, i.e., −1 and 1.

16. In Problem 15, find the value of $\hat{Y}$ in design units at the sixth point on the path of the steepest ascent.

17. Throughout this chapter, we noticed that in almost all experiments discussed, there are some replicated points. Explain why it is useful to replicate some points. What particular information do we gain from such replicates?

18. An experiment involving two factors was designed to be a rotatable CCD. The actual design plan and the results obtained are

Run	x_1	x_2	Yield
1	−1.000	−1.000	23.0
2	1.000	−1.000	24.5
3	−1.000	1.000	19.7
4	1.000	1.000	18.5
5	0.000	0.000	20.6
6	0.000	0.000	22.5
7	−1.414	0.000	17.5
8	1.414	0.000	23.6
9	0.000	−1.414	24.0
10	0.000	1.414	22.4
11	0.000	0.000	22.6
12	0.000	0.000	21.4

(a) Fit a second-order model to these data.
(b) Test the adequacy of the fitted model.
(c) Reduce the fitted model in (a) to its canonical form and describe the nature of the fitted response surface.

19. As a preliminary test to determine the adequacy of the fit of a first-order model, three independent variables were employed at two levels each, forming a 2^3 factorial design. To these 8 points 2 center points were added, giving a total of 10 observations. The actual design plan and results obtained are displayed below. The experiment was conducted in random order.

Run	x_1	x_2	x_3	Yield
1	−1	−1	−1	40
2	1	−1	−1	66
3	−1	1	−1	47
4	1	1	−1	80
5	−1	−1	1	62
6	1	−1	1	60
7	−1	1	1	58
8	1	1	1	64
9	0	0	0	68
10	0	0	0	66

(a) Fit a first-order model to these data.
(b) Construct the ANOVA table for these data and conduct a test for lack of fit.
(c) Test the significance of the regression coefficients.

20. Add the following eight data points to the data in Problem 19.

Run	x_1	x_2	x_3	Yield
1	0	0	0	64
2	0	0	0	59.4
3	−1.682	0	0	67.1
4	1.682	0	0	62.2
5	0	−1.682	0	66.3
6	0	1.682	0	60.5
7	0	0	−1.682	67.3
8	0	0	1.682	63.9

(a) Fit a second-order model to the combined data.
(b) Test the adequacy of the fitted model.
(c) Reduce the fitted model in (a) to its canonical form and describe the nature of the stationary point of the fitted response surface.

21. An experimenter uses a one-half fractional 2^4 factorial design with defining relation $I = ABCD$ to fit the first-order model

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

(a) Construct the X matrix for the proposed model.
(b) Should the design in this problem be considered orthogonal? (A design is called orthogonal if $X'X$ is a diagonal matrix.)
(c) Examine, in the present case, whether the design is variance optimal. Note that the variances of regression coefficients are minimized (i.e., the design is variance optimal) if $X'X = N \times I_k$, where I_k is an identity matrix of order $k \times k$, and N is the number of runs (observations) in the experiment while $k = p + 1$, $p =$ number of coefficients (not including (β_0) in the assumed model).

22. Suppose we add four center points to the one-half fractional 2^4 factorial design with defining relation $I = ABCD$ in Problem 21, and suppose the model we wish to fit is the same as in Problem 21.

(a) Construct the X matrix for the proposed model.
(b) Is the design considered in this problem orthogonal?
(c) Is the design variance optimal?

23. Suppose in Problem 22, instead of adding four center points, we just replicate the one-half fractional 2^4 factorial designs.

(a) Construct the X matrix for the proposed model.
(b) Is the design considered in this problem orthogonal?
(c) Is the design variance optimal?

Chapter 20

STATISTICAL QUALITY CONTROL—PHASE I CONTROL CHARTS

This chapter is not included in the printed copy, but is available for download from the book's website: www.wiley.com/ college/ gupta/ statistics2e

The focus of this chapter is a discussion of phase I control charts for variables, attributes, and learning about process capability.

Topics Covered

- Some valuable tools for achieving quality such as the Pareto chart, cause-and-effect diagram, and defect concentration diagram
- Shewhart $\bar{X}$ and R control chart
- Shewhart $\bar{X}$ and R control chart when process mean μ and process standard deviation σ are known
- Shewhart $\bar{X}$ and S control chart
- Shewhart control chart for individual observations
- Shewhart control chart when sample size is variable
- The p chart with sample size constant
- The p chart with sample size variable
- The np chart
- The c chart
- The u chart
- Process capability

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the difference between variables and attributes.
- Design phase I variable control charts and attribute control charts.
- Set up $\bar{X}$ and R control charts and use them in different scenarios.
- Set up $\bar{X}$ and R control charts when process mean μ and process standard deviation σ are known.
- Apply Western Electric rules to interpret different patterns on $\bar{X}$ and R control charts.
- Set up $\bar{X}$ and S control charts and use them in different scenarios.
- Interpret different patterns on $\bar{X}$ and S control charts.
- Set up control charts and use them for individual values.
- Set up p control charts for the fraction nonconforming and use them in different scenarios.
- Set up np control charts for the number of nonconforming units and use them in different scenarios.
- Set up c *control charts* for the number of nonconformities and use them in different scenarios.
- Set up u control charts for the number of nonconformities per unit and use them in different scenarios.
- Understand the concept of rational subgroups for variable and attribute control charts.
- Estimate the process capability.
- Use statistical packages MINITAB, R, and JMP to set up various control charts in phase I.

Chapter 21

STATISTICAL QUALITY CONTROL—PHASE II CONTROL CHARTS

This chapter is not included in the printed copy, but is available for download from the book's website: www.wiley.com/college/gupta/statistics2e

The focus of this chapter is a discussion of phase II control charts.

Topics Covered

- Basic concepts of the CUSUM control chart
- Designing a CUSUM control chart
- Two-sided CUSUM control charts using numerical procedures
- The fast initial response (FIR) CUSUM control chart
- Combined Shewhart–CUSUM control charts
- CUSUM control chart for controlling process variability
- Moving average (MA) control charts
- Exponentially weighted moving average (EWMA) control charts

Learning Outcomes

After studying this chapter, the reader will be able to

- Understand the difference between phase I and phase II control charts.
- Set up one-sided and two-sided CUSUM control charts.
- Set up one-sided and two-sided tabular CUSUM control charts.
- Understand the advantages of CUSUM control charts over $\bar{X}$ and R control charts.
- Set up moving average (MA) control charts.
- Set up EWMA control charts.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

Appendices

Appendix A: Statistical Tables

1 Summary of Common Probability Distributions*.
2 Table of Binomial Probabilities.
3 Table of Poisson Probabilities.
4 Standard Normal Distribution Table*.
5 The t Distribution Table*.
6 Chi-square Distribution Table*.
7 The F Distribution Table.
8 Values for Constructing Control Charts for Variables.
9 Critical Values for the Sign Test.
10 Critical Values of the Wilcoxon Signed-Rank Test.
11 Quantiles of the Mann-Whitney Test Statistic.
12 Lower and Upper critical values of r in the runs test.
13 Percentage Points of the Studentized Range for 2 through 20 treatments.
14 Transformation of r to z ($z = 0.5 \ln[(1 + r)/(l - r)]$).
15 Critical Values of the Spearman Test Statistic.

Appendix B: Answers to Selected Problems
Appendix C: Bibliography

Because of lack of space Tables 2, 3, and 7 - 15 and Appendix D are not included here. But they are available for downloading on the books' website: www.wiley.com/college/gupta/statistics2e.

*Only these tables are included in the printed copy and the rest of the tables are available at *www.wiley.com/college/gupta/statistics2e*

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

A

STATISTICAL TABLES

Table A.1 Summary of common probability distributions.

Distribution	Probability Function	MGF	Mean	Variance
Discrete uniform	$p(x) = \frac{1}{N}$ $x = 1, 2, \ldots, N$		$\frac{N+1}{2}$	$\frac{N^2-1}{12}$
Hyper-geometric	$h(x) = \dfrac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$ $\mathrm{Max}[0, n-(N-N_1)] \leq x \leq \mathrm{Min}(n, N_1)$		$\frac{nN_1}{N}$	$\left(\frac{N-n}{N-1}\right)\left(\frac{nN_1}{N}\right)\left(1-\frac{N_1}{N}\right)$
Bernoulli	$p(x) = p^x(1-p)^{1-x}$ $x = 0, 1; 0 \leq p \leq 1$	$pe^t + (1-p)$	p	$p(1-p)$
Binomial	$b(x) = \binom{n}{x} p^x(1-p)^{1-x}$ $x = 0, 1, \ldots, n; 0 \leq p \leq 1$	$[pe^t + (1-p)]^n$	np	$np(1-p)$
Poisson	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$	$e^{\lambda(e^t-1)}$	λ	λ
Negative binomial	$p(x) = \binom{x-1}{k-1} p^k(1-p)^{x-k}$ $x = k, k+1, \ldots; 0 \leq p \leq 1$	$\left(\frac{pe^t}{(1-(1-p)e^t)}\right)^k$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
Geometric	$p(x) = p(1-p)^{x-1}$ $x = 1, 2, \ldots; 0 \leq p \leq 1$	$\frac{pe^t}{(1-(1-p)e^t)}$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

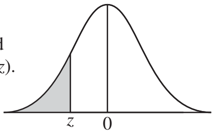
Companion website: www.wiley.com/college/gupta/statistics2e

Table A.1 (*Continued*)

Distribution	Probability Function	MGF	Mean	Variance
Uniform	$f(x) = \frac{1}{b-a} \quad a \leq x \leq b$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ $-\infty < x < \infty,$ $-\infty < \mu < \infty, \sigma > 0$	$e^{\mu t + (\sigma^2 t^2)/2}$	μ	σ^2
Lognormal	$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln x-\mu)^2}$ $x > 0$		$e^{\mu+\sigma^2/2}$	$e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$
Exponential	$f(x) = \lambda e^{-\lambda x}$ $x > 0, \lambda > 0$	$\left(1 - \frac{t}{\lambda}\right)^{-1}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$f(x) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$ $x > 0, \lambda > 0, \nu > 0$	$\left(1 - \frac{t}{\lambda}\right)^{-\nu}$	$\frac{\nu}{\lambda}$	$\frac{\nu}{\lambda^2}$
Weibull	$f(x) = \frac{\beta}{\alpha} \left(\frac{x-\tau}{\alpha}\right)^{\beta-1} e^{-[\frac{x-\tau}{\alpha}]^\beta}$ $\alpha > 0, \beta > 0, \tau > 0$	$\alpha^t \Gamma\left(1 + \frac{t}{\beta}\right)$	$\alpha\Gamma\left(1 + \frac{1}{\beta}\right)$	$\alpha^2\Gamma\left(1 + \frac{2}{\beta}\right) -$ $\left(\Gamma\left(1 + \frac{1}{\beta}\right)\right)^2$
Chi-square	$\frac{1}{2^{n/2}\Gamma(n/2)} w^{n/2-1} e^{-w/2}$ $w \geq 0, n > 0$	$(1-2t)^{-n/2}$	n	$2n$
Student-t	$f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{(n+1)}{2}}$ $-\infty < t < \infty$		0	$\frac{n}{n-2}$

Table A.4 Standard normal distribution table.

The entries in this table give the cumulative area under the standard normal curve to the left of z, $P(Z \le z)$.

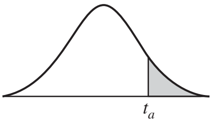


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table A.5 The t distribution table.

The entries in this table give the critical values
of t for the specified number of degrees
of freedom and areas in the righ tail.

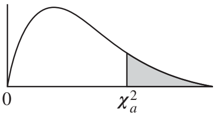


	Area in the right tail under the t distribution curve					
df	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340

36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
41	1.303	1.683	2.020	2.421	2.701	3.301
42	1.302	1.682	2.018	2.418	2.698	3.296
43	1.302	1.681	2.017	2.416	2.695	3.291
44	1.301	1.680	2.015	2.414	2.692	3.286
45	1.301	1.679	2.014	2.412	2.690	3.281
46	1.300	1.679	2.013	2.410	2.687	3.277
47	1.300	1.678	2.012	2.408	2.685	3.273
48	1.299	1.677	2.011	2.407	2.682	3.269
49	1.299	1.677	2.010	2.405	2.680	3.265
50	1.299	1.676	2.009	2.403	2.678	3.261
51	1.298	1.675	2.008	2.402	2.676	3.258
52	1.298	1.675	2.007	2.400	2.674	3.255
53	1.298	1.674	2.006	2.399	2.672	3.251
54	1.297	1.674	2.005	2.397	2.670	3.248
55	1.297	1.673	2.004	2.396	2.668	3.245
56	1.297	1.673	2.003	2.395	2.667	3.242
57	1.297	1.672	2.002	2.394	2.665	3.239
58	1.296	1.672	2.002	2.392	2.663	3.237
59	1.296	1.671	2.001	2.391	2.662	3.234
60	1.296	1.671	2.000	2.390	2.660	3.232
61	1.296	1.670	2.000	2.389	2.659	3.229
62	1.295	1.670	1.999	2.388	2.657	3.227
63	1.295	1.669	1.998	2.387	2.656	3.225
64	1.295	1.669	1.998	2.386	2.655	3.223
65	1.295	1.669	1.997	2.385	2.654	3.220
66	1.295	1.668	1.997	2.384	2.652	3.218
67	1.294	1.668	1.996	2.383	2.651	3.216
68	1.294	1.668	1.995	2.382	2.650	3.214
69	1.294	1.667	1.995	2.382	2.649	3.213
70	1.294	1.667	1.994	2.381	2.648	3.211
71	1.294	1.667	1.994	2.380	2.647	3.209
72	1.293	1.666	1.993	2.379	2.646	3.207
73	1.293	1.666	1.993	2.379	2.645	3.206
74	1.293	1.666	1.993	2.378	2.644	3.204
75	1.293	1.665	1.992	2.377	2.643	3.202
∞	1.282	1.645	1.960	2.326	2.576	3.090

Table A.6 Chi-square distribution table.

The entries in this table give the critical values of χ^2 for the specified number of degrees of freedom and areas in the right tail.



	Area in the right tail under the chi-square distribution curve									
df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.382
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

B

ANSWERS TO SELECTED PROBLEMS

CHAPTER 2

Sections 2.1 and 2.2

3. (a) All the students of the graduation class. (b) All the students in the professor's class. (c) GPA. **5.** (a) Ratio (b) Ratio (c) Ratio (d) Nominal (e) Ratio (f) Ordinal (g) Ratio (h) Ratio (i) Interval (j) Ratio (k) Ratio (l) Ratio (m) Ordinal (n) Nominal.

Section 2.3

1. (b) (1, 2, 3, 4, 5) $\sim$ (24%, 24%, 16%, 22%, 14%) (c) 48%. **3.** (b) (1, 2, 3, 4, 5) $\sim$ (30.56%, 22.22%, 13.89%, 8.33%, 25%) (c) 44.45%.

Section 2.4

3. The line graph does not show any particular pattern. There are some dips and peaks which are occurring randomly. **5.** The stem-and-leaf diagram with increment 5 is more informative. **9.** (b) 13 days. **11.** Ninety percent of the parts have life span between 20 and 48 months. Only one out of 30 parts has a life span more than 50 months. Two out of 30 parts have life spans less than 20 months. **15.** (b) Median consumption of electricity in kilowatt-hours is 248.5 kWh, maximum consumption is 310 kWh, and minimum consumption is 206 kWh.

Section 2.5

1. (a) $\bar{X} = 120.02$, Med$=$ 120.10, mode$=$ 120.1 (b) 1.84 (c) approximately symmetric. **3.** (a) $\bar{X} =$ 22.356, Med$=$ 23.00 (b) 4.73 (c) 100%. **5.** (a) $\bar{X} = 108.47$, Med$=$ 107.50, mode$=$ 100 (b) Range$=$ 20, $S^2 = 50.10$, $S = 7.08$, $CV = 6.53$. **7.** 3.75238. **9.** (51,100–60,100), (46,600–64,600). **11.** (a) $\bar{X} = 22.650, S = 1.461$ (b) (21.189, 24.111), (19.728, 25.572), (18.267, 27.033) (c) 100%, 100%, Chebyshev's inequality is valid.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP,
Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

Section 2.6

1. (a) $\bar{X}_G = 22.47$, $M_G = 23$, Mode$=$ 26.25 (b) $S^2 = 20.165, S = 4.491$. **3.** (a) $\bar{X}_G = 108.8$, $M_G = 108.57$, Mode$=$ 100 (b) $S^2 = 55.138$, $S = 7.425$. **5.** $\bar{X}_G = 49.89$, $S_G = 6.727$, $\bar{X} = 49.56, S = 7.0$.

Sections 2.7 and 2.8

1. (a) $\bar{X} = 12.026$, $S^2 = 0.289$, $S = 0.537$ (b) $Q_1 = 11.753$, $Q_2 = 12.070$, $Q_3 = 12.320$, IQR$=$ 0.567 (c) Outlier 13.60. **3.** Outliers 56, 58, 59. **5.**(I) $\bar{X} = 25.667$, $S = 2.820$, CV $= 10.91\%$, (II) $\bar{X} = 51.194$, $S = 5.966$, CV $= 11.56\%$. Second set has slightly higher variability. **7.** (a) $\bar{X} = 49.56$, $S^2 = 49.00$, $S = 7.00$ (b) $Q_1 = 43$, $Q_2 = 48$, $Q_3 = 57.75$, IQR$=$ 14.75. (c) No outliers.

Section 2.9

1. (a) No correlation (b) 0.075. **3.** (a) Positive correlation (b) 0.822.

Review Practice Problems

1. On the majority of days, 6–10 workers did not come to work. On about 10% of the days, more than 10 workers did not come to work. **3.** Most shifts had one or two machine breakdowns. In about 15% of the shifts, no machine had any breakdown. Very rarely, more than four machines had any breakdown. **5.** Side-by-side bar chart is more informative. **7.** (c) 32.2%. **11.** About 40% of the defects are of type B, whereas type A defects are only about 4%. Defects of types C, D, and E occurred with almost the same frequency. **13.** (d) 46% (e) 54%. **15.** The two frequency distributions are same. **17.** $\bar{X} = 11$, med$=$ 11.50, mode$=$ 7. **19.** $\bar{X} = 420.90$, med$=$ 416.00, mode$=$ 380, 398, 416, 430, 450; slightly right skewed (b) $S^2 = 869.27$, $S = 29.48$, CV $= 7.0\%$, Range$=$ 110.0. **23.** $\bar{X}_G = 33.95, S_G = 7.55, S_G^2 = 56.99, \bar{X} = 33.13, S = 7.51, S^2 = 56.45$. **25.** (a) $\bar{X} = 125.0$, $S^2 = 29.18, S = 5.40$ (b) Empirical rule holds. **27.** (a) $Q_1 = 32$, $Q_2 = 33$, $Q_3 = 35$ (b) IQR$=$ 3 (c) No outliers. **29.** (a) $Q_1 = 30.50$, $Q_2 = 36.00$, $Q_3 = 41.50$, IQR$=$ 11.00 (b) 50% (c) Yes. **31.** (a) $\bar{X} = 779.0, S = 85.6$ (b) 55%. **33.** (a) 68% (b) 100% (c) 2.5% (d) 2.5%. **35.** (a) $\bar{X} = 288.25$, $S = 26.96$ (b) 100% (c) 46.25. **37.** (a) $\bar{X} = 12.733$ (b) $S = 2.840$ (c) CV $= 22.31\%$. **39.** The data in Problem 37 is slightly right skewed, and in Problem 38, it is almost symmetric. **41.** Shipment I: $Q_1 = 60.00$, $Q_2 = 67.50$, $Q_3 = 71.75$, Shipment II: $Q_1 = 42.25$, $Q_2 = 49.50$, $Q_3 = 55.00$ (c) The average number of defective ball bearings per shipment is higher in the first shipment than in the second shipment. However, there is more variability in the second shipment. **43.** 0.314, correlation between inflation rates and interest rates is quite weak. **45.** 0.252, correlation between hours of sleep and test scores is quite weak.

CHAPTER 3

Sections 3.2 and 3.3

1. (a) $A \cup B$ (b) $A \cap B$ (c) $\bar{A} \cap \bar{B}$ or $\overline{(A \cup B)}$ (d) $(A \cap \bar{B}) \cup (\bar{A} \cap B)$ (e) $(\overline{A \cap B})$. **3.** (a) $S =$ {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}, (b) $S =$ {(H, 1) (H, 2) (H, 3) (H, 4)

(H, 5) (H, 6) (T, 1) (T, 2) (T, 3) (T, 4) (T, 5) (T, 6)} (c) $S = \{(1, 1)\ (1, 2) \cdots (1, 6)\ (2, 1)\ (2, 2) \cdots (2, 6) \cdots (6, 6)\}$, (d) $S =$ {BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG}, (e) $S =$ {(HH, 1) (HT, 1) (TH, 1) (TT, 1) (HH, 2) (HT, 2) (TH, 2) (TT, 2) $\cdots$ (HH, 6) (HT, 6) (TH, 6) (TT, 6)}. **5.** $S =$ {MMM, MMC, MME, MCM, MCC, MCE, MEM, MCE, MEE, CMM, CMC, CME, CCM, CCC, CCE, CEM, CCE, CEE, EMM, EMC, EME, ECM, ECC, ECE, EEM, ECE, EEE}, 20/27. **7.** (a) {4} (b) {1, 4, 5, 7} (c) {1, 3, 4, 7} (d) {5} (e) $\emptyset$ (f) $\emptyset$. **9.** (a) $A \cap B \cap C$: The patient is diagnosed with liver cancer and needs a liver transplant and the hospital finds a matching liver in time. (b) $A \cap (B \cup C)$: The patient is diagnosed with liver cancer and the patient needs a liver transplant but the hospital does not find the matching liver in time. OR: The patient is diagnosed with liver cancer and does need a liver transplant but the hospital finds a matching liver in time. OR: The patient is diagnosed with liver cancer who needs a liver transplant and the hospital finds a matching liver in time. (c) $\bar{A} \cap \bar{B} = (\overline{AUB})$: The patient is not diagnosed with liver cancer and does not need a liver transplant. (d) $(\bar{A} \cap \bar{B} \cap \bar{C}) = (\overline{A \cup B \cup C})$: The patient is not diagnosed with liver cancer and does not need a liver transplant and the hospital does not find a matching liver. **11.** {CCCCC, CCCCN, CCCNC, CCNCC, CNCCC, NCCCC, ..., NNNNC, NNNNN}**13.** (a) {HHT, HTH, THH, HHH} (b) {HTT, THT, TTH, TTT} (c) {HHT, HTH, THH} (d) {TTT}.

Section 3.4

1. 20. **3.** (a) 657,720 (b) 810,000. **5.** 480. **7.** 0.02166. **9.** 40. **11.** 48.

Sections 3.5 and 3.6

1. 2/3. **3.** $P(A_1|E) = 0.1569$, $P(A_2|E) = 0.0392$, $P(A_3|E) = 0.2059$, $P(A_4|E) = 0.3529$, $P(A_5|E) = 0.2451$. **5.** (a) 0.2727 (b) 0.4. **7.** 0.4138. **9.** 0.5714.

Review Practice Problems

1. (a) 93.9% (b) 6.1% (c) 95.8% (d) 98.9%. **3.** (a) $E_A \cap E_B = 2$, $E_A \cap \bar{E}_B = 5$, $\bar{E}_A \cap E_B = 3$, $\bar{E}_A \cap \bar{E}_B = 90$, $E_A \cup E_B = 10$, (b) $G \cap G_A \cap G_B = \emptyset$, $G \cap G_A \cap \bar{G}_B = \emptyset$, $G \cap \bar{G}_A \cap G_B = \emptyset$, $\bar{G} \cap \bar{G}_A \cap \bar{G}_B = \emptyset$, $\bar{G} \cap G_A \cap \bar{G}_B = 460$, $G \cap \bar{G}_A \cap \bar{G}_B = 4005$, $\bar{G} \cap \bar{G}_A \cap G_B = 273$, $\bar{G} \cap G_A \cap G_B = 212$. **5.** (a) 0.000119 (b) 0.0005934 (c) 0.0000182. **7.** (b) 0.9999, 0.99999603 (c) $1 - p^2$, $1 - [p^4 + 4p^3(1-p)]$. **9.** 0.368, 0.368, $e^{-1}/k!$. **11.** 0.182, 0.2165, 0.1970. **13.** (a) $1/6^5$ (b) 5/324 (c) 25/108. **15.** (a) 9/24 (b) 15/24. **17.** 0.0456. **21.** 4/13. **23.** 8/9. **25.** 8/39, 1/39, 24/39, 6/39. **27.** (a) 0.2105 (b) 0.2632 (c) 0.5263. **29.** 165/237, 18/29, 45/67. **31.** No, no, yes. **33.** (a) $A = \{3, 4, 5\}$, $B = \{4, 6\}$, $C = \{5, 7, 8\}$ (b) (i) 1/4 (ii) 1/6 (iii) 1/4 (iv) 1/12 (v) 0 (vi) 1 (vii) 1/2 (viii) 1/2.

CHAPTER 4

Sections 4.1 and 4.2

1. (a) 0.17 (b) 0.33 (c) 0.50 (d) 0.48. **3.** (a) Yes (b) No (c) No. **5.** $\mu = 3.75, \sigma^2 = 0.9375$.
7.

$X = x$	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$\mu = 7, \sigma^2 = 5.83$. **9.** (a) 21, 52.47 (b) 19, 23.32.

Sections 4.3 and 4.4

1. $P(X = i) = 1/10, i = 1, 2, \ldots, 10, \mu = 5.5, \sigma^2 = 8.25$. **3.**

X	0	1	2	3	4
$p(x)$	0.0303	0.2424	0.4545	0.2424	0.0303

$\mu = 2, \sigma^2 = 0.7273$. **5.** $S = \{e_i = i | i = 1, 2, \ldots, 20\}$, $P(X = i) = 1/20$, $i = 1, 2, \ldots, 20$, (a) 0.25 (b) 0.45 (c) 0.45. **7.** (a) 0.3576 (b) 0.4551 **9.** (a) 0.4242 (b) 0.5758 (c) 0.8484.

Sections 4.5 and 4.6

1. (a) 0.7238 (b) 0.9601 (c) 0.2762. **3.** $P(X = x) = \binom{n}{x}(p)^x(1-p)^{n-x}, x = 0, 1, \ldots, n$. **5.** (a) 0.000004 (b) 0.189651 (c) 0.998356 **7.** Binomial n =16, $p = 0.75$, 0.99836 **9.** 0.5941.

Section 4.7

1. 0.03762. **3.** 0.00025. **5.** 0.00647. **7.** 0.00028.

Section 4.8

1. (a) 0.7255 (b) 0.7989 (c) 0.0268 (d) 0.2745. **3.** (a) 0.1680 (b) 0.2560 (c) 0.4557. **5.** (a) 0.7127 (b) 0.1336 (c) 0.9858. **7.** 0.4126.

Section 4.9

1. 0.0068. **3.** 0.03195 **5.** 0.8245 **7.** 0.01388.

Review Practice Problems

1. 0.1867, 0.3843, 0.2964, 0.1098, 0.0208, 0.0020, 0.0001, 0.00. **3.** 0.598, 0.402, 0.161. **5.** 0.0036. **7.** (a) $\frac{\binom{13}{x}\binom{39}{13-x}}{\binom{52}{13}}$ (b) $\frac{\binom{13}{y}\binom{39}{13-y}}{\binom{52}{13}}$ (c) $\frac{\binom{13}{x}\binom{13}{y}\binom{39}{13-x-y}}{\binom{52}{13}}$. **9.** 0.000. **11.** (a) 0.3233 (b) 0.0025. **13.** (a) 0.0641 (b) 0.00098. **15.** $\frac{\binom{10}{4}\binom{20}{x-5}}{\binom{30}{x-1}} \times \frac{6}{30 - x + 1}, x = 5, \ldots, 25$. **17.** (a) $\binom{m}{x} p^x(1-p)^{m-x}$ (b) $\binom{n}{y}(1-p)^{my}(1-[1-p]^m)^{n-y}$. **19.** 0.2508. **21.** (a) $(1-p)^{x-1}p$ (b) $\binom{x-1}{k-1}(1-p)^{x-k}p^k$. **23.** 0.1175, $\binom{x-1}{x}(0.85)^4(0.15)^{x-4}$. **29.**

(a) 0.6165 (b) 0.3348 (c) 0.9427 (d) 0.0022. **31.** (a) 0.5155 (b) 0.05 (c) 0.5948. **33.** $\mu = 0.8, \sigma^2 = 0.6691, \sigma = 0.8180$. **35.** (a) 0.8488 (b) 0.9574 (c) 0.5622 (d) 0.0174 (e) 0.9380. **37.** (a) 0.3712 (b) 0.2381 (c) 0.7977 (d) 0.7851. **39.** (a) 0.9997 (b) 0.0210 (c) 0.1259 (d) 0.0355 (e) 0.1275. **41.** 0.5615. **43.** (a) No (b) Yes (c) No. **45.** (a) 20/21 (b) 1/60 (c) 1/15. **47.** (a) $\mu = 4.33,\ \sigma^2 = 2.251$ (b) $\mu = 4.2,\ \sigma^2 = 0.84$. **49.** $\mu = p,\ \sigma^2 = p(1-p)$. **51.** $\mu = [(n+1)/2] + c,\ \sigma^2 = (n^2-1)/12$. **53.** (a) 0.01264 (b) 0.0902 (c) 0.6321. **55.** (a) $p(x) = (1-p)^{x-1}p,\ x = 1, 2, \ldots,\ \mu = 1/p, \sigma^2 = (1-p)/p^2$ (c) 0.01763. **57.** (a) 0.3916 (b) 0.8998 (c) 0.4266.

CHAPTER 5

Sections 5.1 and 5.2

1. (a) 0.0003 (b) 1.0000 (c) 0.0183 (d) 0.00. **3.** $c = 1/48$ (a) 1 (b) 0.7422 (c) 0.5000. **5.** $c = 12$, 0.1808. **7.** $(1 - t/\lambda)^{-1}$, $\mu = 1/\lambda, \sigma^2 = 1/\lambda^2$. **9.** (a) 0.5813 (b) 1 (c) 1.

Section 5.3

1. $\bar{X} = 17.550, S = 1.468$, 80%, 100%, 100%. **3.** $\bar{X} = 10.9, S = 3.463$, 90%, 95%,100%. **5.** 75%.

Section 5.4

1. (a) 2/3 (b) 1/2 (c) 1/2. **3.** $\mu = 7.5,\ \sigma^2 = 18.75,\ \sigma = 4.33$. **5.** 7/12. **7.** $\mu = 3, \sigma^2 = 0.333, \sigma = 0.577$, 1.00.

Section 5.5

1. (a) 0.8164 (b) 0.9082 (c) 0.8413. **3.** (a) 0.6730 (b) 0.8413 (c) 0.9452. **5.** (a) 0.5 (b) 0.0228 (c) 0.1587. **7.** (a) 0.8472 (b) 0.7611 (c) 0.9772.

Section 5.6

1. (a) 0.0548 (b) 0.7118 (c) 0.5670 (d) 0.8790. **3.** (a) 0.5 (b) 0.7933 (c) 0.9282. **5.** $N(32, (3.145)^2)$.

Sections 5.7 and 5.8

1. Normal, $\bar{X} = 27.5, S = 4.662$. **3.** Normal **5.** (a) 0.9630 (b) 0.9596. **7.** (a) 0.9251 (b) 0.4768. **9.** 0.8708.

Section 5.9.1

1. 0.2236. **3.** 8.373, 19.91. **5.** (a) Mean = 54.6, SD = 399.72 (b) 0.0392 (c) 0.0473.

Section 5.9.2

1. (a) 0.2865 (b) 0.2493 (c) 0.6321 (d) 0.7135. **3.** 0.2635. **5.** 0.0000. **7.** (a) 0.2231 (b) 0.1422 (c) 0.7135.

Sections 5.9.3 and 5.9.4

1. 0.0000. **3.** (a) 0.9656 (b) 0.8444 (c) 0.6879. **5.** (a) $\mu = 12$, $\sigma^2 = 720$ (b) 0.2750 (c) 0.7943. **7.** (a) 0.4060 (b) 0.9084 (c) 0.3144. **9.** (a) $\mu = 4$ (b) 0.2057.

Review Practice Problems

1. (a) 0.3085 (b) 0.1587 (c) 1829 (d) 392. **3.** (a) 2.28% (b) 81.75%. **5.** (a) 50.135% (b) $\mu = 0.25$ (c) 13.36%. **7.** Normal. **9.** (a) 0.0052 (b) 17. **11.** (a) 0.0004 (b) 11. **13.** (a) 0.1469 (b) 0.9596. **15.** (a) 0.0036 (b) 63.03. **17.** (a) 1438.49 (b) 0.9772. **19.** (a) 0.9826 (b) 0.8849 (c) 0.9347 (d) 0.0250 (e) 0.0250. **21.** (a) 0.1353 (b) 0.0183 (c) 0.1170 (d) 1. **23.** (a) 0.1199 (b) 0.0710. **25.** (a) 0.4727 (b) 0.3679 (c) 0.2625. **27.** 0.4066. **29.** (a) $\mu = 3$, $\sigma^2 = 3$ (b) $\mu = 4$, $\sigma^2 = 2.67$. **31.** (a) $f(x) = nx^{n-1}$, $0 \le x \le 1$, and 0 elsewhere (b) $2^{-1/n}$ (c) $\mu = n/(n+1), \sigma^2 = n/[(n+1)^2(n+2)]$. **33.** (a) $F(x) = 0$ for $x \le 0$, x^3 for $0 < x < 1$, 1 for $x \ge 1$ (b) 0.088 (c) 0.63 (d) $\mu = 0.75$, $\sigma^2 = 0.0375$. **35.** (a) 2/9 (b) 4/27 (c) 20. **37.**1, 75%; 0.9728, 75%. **39.** (a) 0.7059 (b) 0.4335 (c) 0.7350. **41.** $\mu = 80, \sigma^2 = 1600$. **43.** (a) 0.3679 (b) 0.7769 (c) 0.1448. **45.** (a) (0, 7.20) (b) 0.9835. **47.** (a) 0.4422 (b) 0.7358. **49.** (a) 0.0646 (b) 0.0470. **51.** $\mu = 54.60$, $\sigma^2 = 1.59773 \times 10^5$. **53.** (a) $\mu = 7200$ (b) $\sigma^2 = 3.57696 \times 10^9$. **55.** (a) 0.8867 (b) 0.0285 (c) 0.1005. **57.** (a) 0.2835 (b) 0.1859 (c) 0.3679. **59.** (a) 0.2323 (b) 0.4942 (c) 0.5039. **61.** (a) 0.9951 (b) 0.0274 (c) 0.000. **63.** (a) 1/4 (b) 13/20 (c) 2/5. **65.** (a) $\mu = 19.5, \sigma^2 = 1/12$ (b) 1.00. **67.** $\hat{\mu} = 5.766, \hat{\sigma}^2 = 0.271$. **69.** $\hat{\mu} = 3.132, \hat{\sigma}^2 = 0.691$.

CHAPTER 6

Section 6.2

1. (a) $E(U) = 103$, $V(U) = 689$ (b) $E(U) = 103$, $V(U) = 103$. **3.** $E(T) = 0.80, V(T) = 0.0023$. **5.** $f_1(x_1) = 2(1 - x_1), 0 \le x_1 \le 1$, $f_2(x_2) = 2(1 - x_2)$, $0 \le x_2 \le 1$, $E(X_1) = 1/3$, $E(X_2) = 1/3, Var(X_1) = 1/18, Var(X_2) = 1/18, \rho = -1/2$. **7.** $E(X_1|X_2 = 0) = 1$, $E(X_1|X_2 = 1) = 1$, $Var(X_1|X_2 = 0) = 1, Var(X_1|X_2 = 1) = 0$. **9.** $h_1(u) = e^{-u}$, for $u \ge 0$, and 0 elsewhere.

Sections 6.3 and 6.4

1. $M_u(t) = e^{t^2}, \mu = 0$, $\sigma^2 = 2$. **3.** $M_v(t) = e^{(13/2)t^2}, \mu = 0$, $\sigma^2 = 13$. **5.** $M_u(t) = (1 - t)^{-1}$.

Review Practice Problems

1. $M_Y(t) = (1 - t/\lambda)^{-\Sigma\gamma_i}$. **3.** $M_Y(t) = e^{(\Sigma\lambda_i)(e^t - 1)}$. **5.** $\mu_T = 350$, $\sigma_T = 3.606$. **7.** $\mu_T = 0.0450$, $\sigma_T = 0.0009$. **9.** (a) $X_2 \sim Bin(n, p_2)$ (b) $E(X_2) = np_2$, $Var(X_2) = np_2(1 - p_2)$

(c) $-np_1p_2$. **11.** $\mu_T = 4.5n, \sigma_T^2 = 8.25n$. **13.** 2341. **15.** 1,112,430. **17.** $p(0) = \frac{9}{24}$, $p(1) = \frac{8}{24}$, $p(2) = \frac{6}{24}$, $p(3) = 0$, $p(4) = \frac{1}{24}$, $\mu = 1, \sigma^2 = 1$. **25.** (a) $\mu_1 = 2$, $\mu_2 = 3$, $\sigma_1 = \sigma_2 = 1$, $\rho = 4/5$ (b) $c = \frac{5}{6\pi}$ (c) $f_1(x) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)(x-2)^2}, -\infty < x < \infty$, $f_2(y) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)(y-3)^2}, -\infty < y < \infty$.

CHAPTER 7

Section 7.1

1. (a) All employees of the manufacturing company (b) All the chips manufactured in that batch (c) All the voters in that metropolitan area. **3.** \$ 396,000, 39.38×10^8. **5.** 32.20, 5.25.

Section 7.2

1. Approximately $N(28, (1.5)^2)$. **3.** (a) Decreases from $\frac{\sigma}{6}$ to $\frac{\sigma}{8}$ (b) Decreases from $\frac{\sigma}{10}$ to $\frac{\sigma}{20}$ (c) Decreases from $\frac{\sigma}{9}$ to $\frac{\sigma}{18}$ (d) Decreases from $\frac{\sigma}{16}$ to $\frac{\sigma}{24}$. **5.** (a) 0.1587 (b) 0.5 (c) 0.8904. **7.** (a) Approximately $N(0.5, (0.05)^2)$ (b) $P(\hat{p} > 0.60) = P\left(\frac{\hat{p}-0.5}{0.05}\right) = P(Z > 2.00) = 0.0228$.

Section 7.3

1. (a) 0.05 (b) 0.975 (c) 0.025 (d) 0.95 (e) 0.05. **3.** (a) 3.58 (b) 5.06 (c) 3.22 (d) 3.53. **7.** (a) 20.483 (b) 34.170 (c) 40.646.

Section 7.4

1. (a) $g(x_{(n)}) = nx_{(n)}^{n-1}$ for $0 < x_{(n)} < 1$, and 0 elsewhere (b) $g(x_{(1)}) = n(1 - x_{(1)})^{n-1}$ for $0 < x_{(1)} < 1$, and 0 elsewhere (c) $g(x_{(r)}) = \frac{n!}{(r-1)!(n-r)!}x_{(r)}^{r-1}(1 - x_{(r)})^{n-r}$ for $0 < x_{(r)} < 1$, and 0 elsewhere. **3.** $e^{-100\lambda y}$. **5.** (a) $g(y) = (n/15^n)(y - 15)^{n-1}$ (b) $15(2n + 1)/(n + 1)$. **7.** (a) $g(x_{(11)}) = (2.586584 \times 10^5)\left(\frac{x-15}{15}\right)^{10}\left(\frac{30-x}{15}\right)^{10}$ (b) 29.32.

Review Practice Problems

1. (a) 0.1151 (b) 0.0013 (c) 0.9559. **3.** (a) Approximately $N(0.8,\ (0.0179)^2)$ (b) 0.9974. **5.** (a) $[1 - F(x)]^{100}$ (b) $100[1 - F(x)]^{99}f(x)dx$. **7.** (a) 0.9544 (b) 0.7745. **15.** $S^2 \sim 7.2\chi^2_{20}$, $\mu = 144$, $\sigma^2 = 2073.6$. **17.** $g(t) = n\lambda e^{-n\lambda t}$, 0.0498. **19.** 0.0012, 0.1816.

CHAPTER 8

Section 8.2

1. $\hat{p} = \bar{X}$. **3.** (b) $V(\hat{\mu}_1) = 9\sigma^2$, $V(\hat{\mu}_2) = 17\sigma^2$, $\hat{\mu}_1$ is a better estimator. **5.** (a) 9.850 (b) 1.7271. **7.** $\hat{\lambda} = \bar{X}$. **9.** $\hat{\theta} = \text{Max}(X_1, X_2, \ldots, X_n) = X_{(n)}$.

Section 8.3

1. (a) (45.562, 48.049) (b) (45.762, ∞), (0, 47.849). **3.** (a) (8.164, 9.036) (b) (8.260, ∞), $(0, 8.940)$. **5.** (8.608, 9.378). **7.** (16.109, 20.891). **9.** $(0, 87.605), (85.130, \infty)$. **11.** (1301.7, 1513.3); we may conclude with 99% confidence that $\mu = 1400$.

Section 8.4

1. $(-19.48, 6.48)$. **3.** $(-80.4, 353.8)$, yes. **5.** $(-0.071, 0.711)$, no. **7.** $(-0.852, -0.302)$. **9.** $(-22.74, -15.36)$.

Sections 8.5 and 8.6

1. (43.257, 248.496). **3.** (a) (1.081, 1.839) (b) (– 0.1888, 0.0088). **5.** ($15.2423 \leq \sigma^2 \leq 48.3828$), ($3.9041 \leq \sigma \leq 6.9557$), 4.0591, 6.5823. **7.** (2.62, 12.16), ($2.35 \leq \sigma_1^2 \leq \infty$), ($0 \leq \sigma_2^2 \leq 14.69$).

Section 8.7

1. (0.3571, 0.4429). **3.** (–0.1836, 0.0036). **5.** (–0.0597, –0.0063). **7.** (–0.0217, 0.0484). **9.** (0.3640, 0.4360).

Section 8.8

1. 49. **3.** (a) 1201 (b) 1568 (c) 2135. **5.** (a) 1201 (b) 1145. **7.** 1509. **9.** 39.

Review Practice Problems

1. $(5.152\sigma/L)^2$. **3.** (a) $(7.867, 7.973)$ (b) $(0.0103, 0.0681)$. **5.** (a) $(-4.80, 5.30)$ (b) (3.994, 12.296). **7.** (25.67, 72.80). **9.** (–1.285, 2.625). **11.** (–16.15, 3.39). **13.** (–0.235, 10.715). **15.** $(0.0486, 11.5964)$. **17.** $(513.059, 518.681)$. **19.** $(67.516, 68.584)$. **21.** (a) (0.422, 0.867) (b) $(-0.006, 0.806)$. **23.** (a) (0.2637, 0.7491) (b) (–3.8568, –0.1432). **25.** (a) $(0.4319, 1.7451)$ (b) $(1.6088, 2.0892)$. **27.** (0.3521, 0.8479). **29.** $\hat{\mu} = \bar{X}$, yes, $\bar{X} \sim N(\mu, \sigma_0^2/n)$. **31.** ±0.49. **33.** (11.8069, 12.1931). **35.** (–13661.2523, –6331.2523). **37.** (a) (–9.573, 4.427) (b) $(-9.704, \infty)$, $(-\infty, -4.304)$. **39.** $(-0.2428, -0.0172)$. **41.** 31. **43.** (0, 1.8857). **45.** (0.3436, 0.4564). **47.** (–0.2130, 0.0330). **49.** (–0.0681, 0.0021), width of the confidence interval has increased.

CHAPTER 9

Section 9.2

1. H_0: $\mu = 4$ versus H_1: $\mu > 4, Z = 2.4,$ reject H_0. **3.** Type II error smaller, $\beta = 0.00$. **5.** $n = 54$.

Section 9.3

1. $Z = 4$, reject H_0

μ_1	800	802	804	806	808	810	812	814
$\gamma(\mu_1)$	0.01	0.05	0.16	0.37	0.63	0.84	0.95	0.99

The power curve is obtained by plotting μ_1 versus $\gamma(\mu_1)$. **3.** $Z = -5.00$, reject $H_0, 1 - \beta = 0.8431$. **5.** $(\bar{X} < 0.2488$ or $\bar{X} > 0.2512)$; $1 - \beta = 0.3203$.

Section 9.4

1. p-value $= 0.349$, do not reject H_0. **3.** p-value $= 0.016$, do not reject H_0. **5.** p-value $= 0.911$, do not reject H_0. **7.** p-value $= 0.025$, reject H_0.

Section 9.5

1. $Z = -2.052$, reject H_0. **3.** $T = 2.706$, reject H_0. **5.** $Z = -0.77$, do not reject H_0, p-value 0.780.

Section 9.6

1. $Z = -1.7003$, reject H_0. **3.** $Z = 0.6187$, do not reject H_0, 0.0645. **5.** $Z = -8.576$, p-value is 0.00, reject H_0.

Section 9.7

1. $Z = 1.6$; do not reject H_0. **3.** $t = -0.37$, do not reject H_0, p-value$=$ 0.715. **5.** $t = 3.96$, reject H_0, p-value 0.000. **7.** $t = -0.37$, do not reject H_0, p-value 0.716.

Section 9.8

3. $Z = 1.84$, do not reject H_0, p-value 0.066. **5.** $Z = -5.06$, reject H_0, p-value 0.00. **7.** $Z = -0.7924$, p-values (a) 0.214 (b) 0.786 (c) 0.428, in either case do not reject H_0.

Section 9.9

1. $\chi^2 = 15.68 > \chi^2_{8;0.05}$, reject H_0. **3.** $\chi^2 = 6.8 < \chi^2_{4;0.05}$, do not reject H_0. **5.** $\chi^2 = 0.0037$, reject H_0.

Section 9.10

1. (a) $F = 1.108$, do not reject H_0 (b) $t = 1.372$, do not reject H_0. **3.** $F = 1.3972$, do not reject H_0. **5.** $F = 0.4868$, do not reject H_0.

Sections 9.11 and 9.12

1. (a) (624.46, 635.54) (b) reject H_0. **3.** (a) (17.015, 21.785) (b) do not reject H_0. **5.** (a) $(-0.332, 1.092)$ (b) do not reject H_0. **7.** (a) (0.4569, 0.7431) (b) do not reject H_0.

Review Practice Problems

1. (a) $Z = -3$, reject H_0 (b) $\gamma(\mu_1) = 0.5675$ (c)

μ_1	1300	1325	1350	1375	1400	1425	1450	1475
$\gamma(\mu_1)$	0.996	0.98	0.92	0.79	0.57	0.33	0.14	0.04

The power curve is obtained by plotting μ_1 versus $\gamma(\mu_1)$. **5.** Reject H_0 if $\bar{X} < 39.0229$ or $\bar{X} > 40.9771$.

μ_1	38	38.2	38.4	38.6	38.8	39	39.2	39.4	39.6	39.8	40	40.2
$\gamma(\mu_1)$	0.996	0.985	0.95	0.87	0.72	0.52	0.32	0.16	0.06	0.01	0.02	0.06

The power curve is obtained by plotting μ_1 versus $\gamma(\mu_1)$. **7.** $t = 1.395$, do not reject H_0. **9.** $t = 1.786$, do not reject H_0. **11.** $\chi^2 = 29.4$, reject H_0. **13.** (a) $F = 1$, do not reject H_0 (b) $t = 28.57$, reject H_0. **15.** (a) $F=$ 0.666, do not reject H_0 (b) Reject H_0. **17.** (a) $F=$ 0.514, do not reject H_0 (b) $t = 4.859$, reject H_0. **19.** (a) $F=$ 0.931, do not reject H_0 (b) $t = 0.9005$, do not reject H_0. **21.** (a) $F=$ 1.2958, do not reject H_0 (b) $Z = 2.883$, reject H_0. **23.** (a) $F=$ 1.7602, do not reject H_0 (b) $Z = 3.341$, reject H_0. **25.** $t = 4.00$, reject H_0. **27.** $Z = 21.86$, reject H_0. **29.** $F = 0.6251$, do not reject H_0. **31.** $F = 1.410$, do not reject H_0. **33.** $F = 0.897$, do not reject H_0. **37.** $Z = 2.5$, reject H_0. **39.** (a) p-value$=$ 0.00, reject H_0 (b) p-value$=$ 0.00, reject H_0. **41.** $Z = 1.33$, do not reject H_0, p-value$=$ 0.1836, power$=$ 0.9147. **43.** (a) $Z = -9.8918$, reject H_0 (b) $\beta = 0.0122$, $1 - \beta = 0.9878$. **45.** (a) $Z = 2.86$, reject H_0 (b) $\beta = 0.0004$, $1 - \beta = 0.9996$. **47.** $t = -0.99$, do not reject H_0, p-value$=$ 0.1689. **49.** $t = -1.34$, do not reject H_0, $p -$ value > 0.20. **51.** (a) H_0: $\mu \geq 16$ versus H_1: $\mu < 16$ (b) $t = 0.0992$, do not reject H_0 (c) 0.539. **53.** Reject H_0, p-value $<$ 0.005. **55.** (a) H_0: $p = 0.5$ versus H_1: $p \neq 0.5$ (b) Reject H_0, p-value $= 0.00, Z = -4.00$. **57.** $Z = 1.126$, do not reject H_0, p-value$=$ 0.1301. **59.** $\chi^2 = 12.446$, do not reject H_0. **61.** $F = 0.8335$, do not reject H_0. **63.** $F = 0.3335$, do not reject H_0. **65.** (0.3571, 0.4429), reject H_0. **67.** $Z = -1.775$, p-value$=$ 0.076, do not reject H_0. **69.** $Z = -2.49$, p-value$=$ 0.0128, reject H_0.

CHAPTER 10

Section 10.1

1. 0.7165; 0.0333. **3.** 0.5481; 0.0002. **5.** $f(t) = \frac{\alpha\beta t^{\beta-1}}{(1+\alpha t^\beta)^2}, H(t) = \ln(1 + \alpha t^\beta), R(t) = \frac{1}{1+\alpha t^\beta}$.

Section 10.2

1. $\hat{\mu} = 2353.33$. **3.** (0.3813, 0.8339). **5.** $\hat{\lambda} = 0.0069, (0.0015, 0.0173)$.

Sections 10.3 and 10.4

1. (a) Sampling continues if $-530 + 91.95m \leq T_m \leq 530 + 91.95m$, reject H_0 if $T_m < -530 + 91.95m$, do not reject H_0 if $T_m > 530 + 91.95m$ (b) Sampling continues if $-520 + 91.95m \leq T_m \leq 405 + 91.95m$, reject if $T_m < -520 + 91.95m$, do not reject H_0 if $T_m > 405 + 91.95m$. **3.** $\hat{\mu} = 8563.57$, 95% CI$=$ (4590.14, 21299.76). **5.** (a) $\hat{\mu} = 8700.71$ (b) $\hat{R}(8000) = 0.3987$ (c) 1.1493×10^{-4} (d) $(4663.65, 21640.79)$ (e) $(4.6209 \times 10^{-5}, 2.14424 \times 10^{-4})$ (f) 5142.96. **9.** $\hat{\mu} = 4.1080, \hat{\sigma}^2 = 0.02592$; estimate of mean time to failure$=$ 61.618, estimate of variance of time to failure$=$ 99.7.

Review Practice Problems

1. MTBF $= 9491$ hours, $h(t) = 1.054 \times 10^{-4}$ **3.** $\hat{\lambda} = \frac{k}{T_k}, T_k = \sum_{i=1}^{k} t_{(i)} + (n-k)t_{(k)}$. **5.** $\hat{R}(4000) = 0.615$. **7.** (a) (1733, 10931), 1939 (b) 0.569; 0.356 (c) 2460. **9.** (a) Sampling continues if $-43356 + 2735m \leq T_m \leq 33769 + 2735m$, reject H_0 if $T_m < -43356 + 2735m$, do not reject H_0 if $T_m > 33769 + 2735m$.

CHAPTER 11

Review Practice Problems

1. See Section 11.2. **3.** Tabular and graphical displays. See Section 11.4. **5.** See Section 11.5. **7.** 2D scatter plot shows that variable "Fare" tends to increase with the "Age." Passengers who pay high "Fare" (>100) seems to survive. 3D scatter plot also shows that the variable "Fare" tends to increase with the "Age," but the "Fare" seems to be lower among the people who died than who survived. Specially, the middle aged who paid a high fare seem to survive better than the middle aged who paid a lower fare. **9.** (a) All females were survived, while all males were died. (b) Survival rate for class 1 passengers is somewhat higher than the other two groups. **11.** Class 0 error rate$=$ 28.57%, class 1 error rate$=$ 0%. **13.** (b) When cutoff value increases, class 0 error rate decreases and class 1 error rate increases. Cutoff value close to 0.46. **15.** (a) Abdomen and Wrist are the significant classifiers. (b) Class 0 error rate$=$ 5.88%, class 1 error rate$=$ 13.17%, and overall error rate$=$ 10.71%. (c) Variables Abdomen, Height, and Wrist are the significant classifiers. (d) Regression method, suitable. **17.** Abdomen, Knee, Height, and Wrist are the significant classifiers. (a) A positive linear trend. (b) Correlation$=$ 0.70. **19.** All the "D" grades and one "C" grade are classified as "D" grades. Rest of the grades are classified as "B" grades indicating a binary classification.

CHAPTER 12

Review Practice Problems

1. (a) 37.177 (b) 63.5 (c) 0.995. **3.** (c) Both Euclidean and city-block distances provide the same key information about distances between items as such the graphs convey similar information though those are not exactly the same. **5.** Patient pairs "P011" and "P015" and "P014" and "P015" show perfect matching rates while patient pairs "P012" and "P013" and "P013" and "P015" show zero matching rate. **7.** (b) (S1, S3, S4), (S5, S6, S7), (S9, S10), and (S2). The students within those clusters seem to be statistically closer. **9.** (b) (S1, S4, S3, S8), (S5, S7, S6), (S9, S10), and S2. **11.** (a) (S1, S3, S4, S8), (2.95, 3.25); (S5, S6, S7), (3.70, 3.00); (S9, S10), (3.15, 3.85); (S2), (4.00, 3.90). (b) The clusters (S1, S3, S4, S8), (S9, S10) group the students who moderately and highly improved their GPA at the college compare to high school, respectively but the cluster (S5, S6, S7) groups the students who have that of the opposite behavior. The student S2 seems to maintain the GPA without any considerable fluctuation. **13.** (a) (S1, S3, S4, S8), (S5, S6), (S2, S9, S10), (S7). **15.** (a) (A, B, C); (D, E, F, J); (G, H, I) (b) Final clusters: (A, B, C), (1, 0.33); (D, E, F, J), (4.00, 7.00); (G, H, I), (6, 1.33). **21.** (a) BIC values attain their maximums at three clusters. (b) The BIC values of "EII," "VII," "EEE," and "VEE" are very close to each other when the number of clusters is three. Favor spherical shape clusters with or without the same volumes.

CHAPTER 13

Section 13.2

1. Do not reject H_0, p-value$=$ 0.395. **3.** Reject H_0, p-value$=$ 0.000. **5.** Do not reject H_0, p-value$=$ 0.332. **7.** Reject H_0, p-value$=$ 0.000. **9.** Reject H_0, p-value$=$ 0.0108.

Section 13.3

1. Do not reject H_0, p-value$=$ 0.177. **3.** Reject H_0, p-value$=$ 0.000. **5.** Reject H_0, p-value$=$ 0.000.

Section 13.4

1. Do not reject H_0, p-value$=$ 0.187. **3.** Reject H_0, p-value$=$ 0.000. **5.** Reject H_0, p-value$=$ 0.000.

Review Practice Problems

1. Reject H_0, p-value$=$ 0.0162. **3.** $\chi^2 = 6.0808$, reject H_0. **5.** $\chi^2 = 19.30$, do not reject H_0. **7.** $\chi^2 = 6.1393$, reject H_0. **9.** $\chi^2 = 5.1625$, do not reject H_0. **11.** $\chi^2 = 3.111$, do not reject H_0. **13.** Do not reject H_0, p-value$=$ 0.843. **15.** Do not reject H_0, p-value$=$ 0.1356. **17.** Do not reject H_0, p-value$=$ 0.406. **19.** Reject H_0, p-value$=$ 0.003. **21.** Reject H_0, p-value$=$ 0.008. **23.** Do not reject H_0, p-value$=$ 0.333.

CHAPTER 14

Section 14.2

1. Do not reject H_0, p-value $=$ 0.7539. **3.** Do not reject H_0, p-value $=$ 0.8811. **5.** Do not reject H_0, p-value $=$ 0.936. **7.** Do not reject H_0, p-value $=$ 0.5034. **9.** Reject H_0, p-value $=$ 0.040.

Section 14.3

1. Do not reject H_0, p-value $=$ 0.1250. **3.** Reject H_0, p-value $=$ 0.0223. **5.** Do not reject H_0, p-value $=$ 0.3984.

Section 14.4

1. Do not reject H_0. **3.** Do not reject H_0. **5.** Do not reject H_0. **7.** Reject H_0.

Section 14.5

1. $r_s = -0.3030$, do not reject H_0. **3.** $r_s = -0.1643$, do not reject H_0. **5.** $r_s = 0.4703$, do not reject H_0. **7.** $r_s = -0.4212$, do not reject H_0.

Review Practice Problems

1. Do not reject H_0, p-value $=$ 1.0. **3.** $r = 15$, reject H_0. **5.** $Z = -2.513$, reject H_0. **7.** $r = 16$, do not reject H_0, p-value $=$ 0.210. **9.** $Z = -2.412$, reject H_0. **11.** Do not reject H_0. **13.** $r = 2$, reject H_0. **15.** $r_s = 0.2815$, do not reject H_0.

CHAPTER 15

Section 15.2

1. (a) No (b) $\hat{Y} = 135.7 + 0.237X$ (c) $\hat{\beta}_0 = 135.70$, $\hat{\beta}_1 = 0.2371$. **3.** (a) No (b) $\hat{Y} = 105.5 - 0.11X$ (c) $\hat{\beta}_0 = 105.5$, $\hat{\beta}_1 = -0.11$. **5.** (a) Yes (b) $\hat{Y} = -2.53 + 0.0333X$ (c) $\hat{\beta}_0 = -2.53, \hat{\beta}_1 = 0.0333$. **7.** (a) No (b) $\hat{Y} = 55.5 + 0.347X$ (c) $\hat{\beta}_0 = 55.5, \hat{\beta}_1 = 0.347$. **9.** (a) Yes (b) $\hat{Y} = 29.83 + 4.829X$ (c) $\hat{\beta}_0 = 29.829$, $\hat{\beta}_1 = 4.829$. **11.** (a) Yes (b) $\hat{Y} = 18.8 + 0.926X$ (c) $\hat{\beta}_0 = 18.80, \hat{\beta}_1 = 0.926$.

Sections 15.3 and 15.4

1. (a) 2.36 (b) $(8.2257, 10.3205)$, $(1.1057, 1.7673)$ (c) (12.737, 15.863) (d) (10.490, 18.110). **3.** (a) 78.97 (b) (−59.9875, 97.5875), (0.3734, 1.4786) (c) (148.283, 156.921) (d) (133.439, 171.765). **5.** (a) 28.71 (b) (59.1546, 112.0854), (−0.7904, 0.6324) (c) (79.0074, 86.7188) (d) (70.1436, 95.5826). **7.** (a) 39.89 (b) (7.2068, 82.9932), (−0.2218, 2.3438) (c) (74.38, 83.66) (d) (64.36, 93.68). **9.** (a) 28.21 (b) (−31.9013, 33.4013), (−1.7879, 4.2879) (c) (7.61, 17.64) (d) (0, 26.55). **11.** (a) 15.8 (b) (-16.7760, 25.3760), (4.2903, 7.1097) (c) (91.64, 99.36) (d) (85.05, 105.95).

Section 15.5

1. Do not reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$. **3.** Do not reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$. **5.** Do not reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$. **7.** Do not reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$. **9.** Reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$. **11.** Do not reject H_0: $\beta_0 = 0$, reject H_0: $\beta_1 = 0$.

Section 15.6

1. (a) Provides a good fit (b) F-statistic$=$ 96.18, p-value is 0.000. **3.** (b) F-statistic$=$ 12.39, provides a good fit (c) p-value is 0.002 (d) $R^2 = 40.8\%$. **5.** (b) F-statistic$=$ 0.40, does not provide a good fit (c) p-value is 0.544 (d) $R^2 = 4.78\%$. **7.** (b) F-statistic$=$ 0.000, does not provide a good fit (c) p-value is 0.978 (d) $R^2 = 0.008\%$. **9.** (b) F-statistic$=$ 97.88, provides a good fit (c) p-value is 0.000 (d) $R^2 = 94.22\%$.

Section 15.7

1. (a) Normality assumption valid (b) Model assumptions satisfactory. **3.** (a) Normality assumption valid (b) Model assumptions satisfactory. **5.** (a) Normality assumption valid (b) Model assumptions satisfactory. **7.** (a) Normality assumption valid (b) Model assumptions satisfactory. **9.** (a) Normality assumption valid (b) Model assumptions satisfactory. **11.** (a) Normality assumption valid (b) Model assumptions satisfactory.

Section 15.8

1. (a) $\log(\hat{Y}) = 1.71 + 0.00579X$ (b) $\hat{\beta}_0 = 1.71, \hat{\beta}_1 = 0.00579$ (c) Reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$. **3.** (a) $\log(\hat{Y}) = 1.369 + 0.349\ \log(X)$ (b) $\hat{\beta}_0 = 1.369, \hat{\beta}_1 = 0.349$ (c) reject H_0: $\beta_0 = 0, H_0$: $\beta_1 = 0$.

Section 15.9

1. Do not reject H_0: $\rho = 0$. **3.** Reject H_0: $\rho = 0$. **5.** Reject H_0: $\rho = 0$.

Review Practice Problems

3. ($\hat{\beta}_0 = 48.00$, $\hat{\beta}_1 = -0.3167$, $\hat{Y} = 48.00 - 0.3167X$, (39.96, 56.04), ($-0.442, -0.191$), $\hat{Y}_0 \pm (2.228)\sqrt{((1/12) + (x_0 - 60)^2/6000)(19.033)}$. **5.** $\hat{\beta}_0 = 6.2825$, $\hat{\beta}_1 = 0.1831$, $\hat{\eta} = 6.2825 + 0.1831X$, $(6.1415, 6.4245)$, $(0.1757,\ 0.1903)$, $\hat{Y}_0 \pm (2.776)\sqrt{((1/6) + (x_0 - 17.5)^2/437.50)(0.003)}$. **7.** $\hat{\beta}_0 = 2.7$, $\hat{\beta}_1 = 0.193$, $\hat{Y} = 2.7 + 0.193X$, $(2.197, 3.203)$, $(0.127, 0.258)$ $\hat{Y}_0 \pm (2.776)\sqrt{((1/6) + (x_0 - 7)^2/70)(0.037857)}$

$X = x_0$	2	4	6	8	10	12
$\hat{Y}_0$	3.0857	3.4714	3.8571	4.2429	4.6286	5.0143
95% confidence band	±0.3909	±0.2934	±0.2298	±0.2298	±0.2934	±0.3909

9. 12.917, (10.075, 15.759). **11.** Reject H_0: $\rho = 0$, $(0.9217, 0.9915)$. **13.** Reject H_0: $\rho = 0$, $(0.9839, 0.9992)$. **15.** (a) $\hat{Y} = 8.0250 + 0.3715X$ (b) $\hat{Y} = 18.1453 + 2.4229X$ (c) Assumptions are such that the independent variable is measured without error, while the dependent variable has a random error component. **17.** (a) $\hat{Y} = -1.6938 + 2.3643X$ (b) 69.2356 (c) 40.897. **19.** $(29.5195, 30.9227)$. **21.** No, $X = 11$ is outside the experimental range. **23.** Residual plots suggest some abnormalities, quadratic model $Y = 18.38 + 764.7X - 175.6X^2$ is a better fit. **25.** Do not reject H_0: $\rho = 0$. **27.** (a) Linear model (b) $\hat{Y} = -19.7 + 0.322X$ (c) Model assumptions not violated. **29.** (a) $\hat{Y} = 31.4 + 0.017X$, $\hat{Y} = 5.60 + 3.05X - 0.0677X^2$ (b) Quadratic model is a better fit. **31.** (a) $\hat{Y} = 58.3 + 0.433X$, some departure from the assumptions (b) $(51.900, 64.754)$, (0.221, 0.645) (c) (54.40, 84.77), (57.79, 88.30), (61.39, 92.49). **33.** (a) $r = -0.872$ (b) Do not reject H_0: $\rho = 0$, p-value= 0.972 $>\alpha$. **35.** (a) $\hat{Y} = 129.787 - 24.02X$ (c) Model is a good fit, F-ratio= 352.27. **37.** (a) $\log(\hat{Y}) = 2.118 - 0.0984X$ (b) 93.334, 93.757 using model of Problem 35.

CHAPTER 16

Section 16.3

1. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, i = 1, 2, \ldots n, E(\varepsilon) = 0, V(\varepsilon) = \sigma^2$. **5.** (a) $\hat{\beta}_0 = -63.5645, \hat{\beta}_1 = 1.0742, \hat{\beta}_2 = 1.0153$ (b) 88.0716. **7.** $\hat{Y} = 7.46 - 0.030X_2 + 0.521X_3 - 0.102X_4 - 2.16X_5$ (b) Model is significant, p-value= 0.002 (c) (6.848, 11.143). **9.** Do not reject H_0, p-value $= 0.1596$. **11.** (a) $\hat{Y} = 158 + 15.5X_1 - 0.911X_2$ (b) Reject H_0, p-value is 0.001.

Section 16.4

1. $\hat{Y} = 127.4 - 1.460X_1 - 5.98$Lab, indicator variables 1 for Lab A and 0 for Lab B. **3.** For private: $\hat{Y} = 21.80 + 0.039X1$, for public $Y = 28.14 + 0.039X_1$. **5.** Brick: $\hat{Y} = -18.3568 + 2.84465X_1 + 28.6835X_2 - 2.41353X_3 + 25.8673X_4$.
Stucco: $\hat{Y} = -27.8076 + 2.84465X_1 + 28.6835X_2 - 2.41353X_3 + 25.8673X_4$.
Vinyl: $\hat{Y} = -5.33475 + 2.84465X_1 + 28.6835X_2 - 2.41353X_3 + 25.8673X_4$.
Wood: $\hat{Y} = -31.013 + 2.84465X_1 + 28.6835X_2 - 2.41353X_3 + 25.8673X_4$.
7. $\hat{Y} = -27.84 + 4.6X_1 + 20.5X_2 + 24.1X_4$. **9.** Fitted model is significant, p-value= 0.00.

Sections 16.6 and 16.7

1. (a) $\hat{Y} = 9.21 + 1.192X_4 - 0.318X_5 - 2.609X_6 - 1.094X_7$ (b) Same as in part (a). **3.** $\hat{Y} = 9.144 - 0.2X_1 + 0.09X_2 + 1.14X_4 - 0.33X_5 - 2.52X_6 - X_7$, $C_p = 6.1$, PRESS =

$40.0492, R^2_{pred} = 74.1$, Model in Problem 1 is better. **5.** (a) $\hat{Y} = -2.14 + 0.513X_1 - 0.03052X_3 - 0.0250X_1^2 + 0.00433X_2^2$ (b) The model has smaller bias due to sampling error. **7.** (a) $\hat{Y} = 4.66 + 0.511X_3 - 0.1242X_4$ (b) Both models are the same.

Section 16.8

1. $\hat{\eta} = -22.2 + 0.707X_1$. **3.** $\hat{\eta} = -1.59 + 0.118X_1$.

Review Practice Problems

1. $\hat{\beta} = (X'X)^{-1}X'Y$. **5.** (a) $\hat{Y} =$ 0.41 + 0.715 X_1 + 0.485 X_3 − 0.137 X_4 + 1.07 X_6 (b) $\hat{\sigma}^2 = 0.7456$. **7.** $\begin{pmatrix} 0.024713 & 0.001563 \\ 0.001563 & 0.003163 \end{pmatrix}$ (0.7456). **9.**

Source	*DF*	*SS*	*MS*	*F*-ratio	*P*-value
Regression	4	22.7821	5.6955	7.64	0.002
Residual error	14	10.4389	0.7456		
Total	18	33.2211			

$R^2 = 68.6\%, R^2_{adj} = 59.6\%$.

11. (a) Do not reject H_0 for $i = 2$, 4, and 5 (b) (-0.0297 ± 0.56478), (0.5205 ± 0.29151), (-0.10180 ± 0.11452), (-2.161 ± 5.13727). **13.** $R^2 = 71.1\%$, $R^2_{adj} = 56.7\%$. **15.** No, since predictor values fall outside the experimental region. **17.** $\hat{Y} = 18.0 + 1.50X_1 + 0.877X_2$. **19.** (a) $\hat{Y} = -1.77 + 0.421X_1 + 0.222X_2 - 0.128X_3 - 0.0193X_1^2 - 0.0074X_2^2 + 0.00082X_3^2 - 0.0199X_1X_2 + 0.00915X_1X_3 + 0.00258X_2X_3$ (b) p-value= 0.00 (c) Model assumptions valid (d) Do not reject H_0, p-value= 0.2076. **21.** $\hat{Y} = -0.259 + 0.0782X_1 + 0.121X_2 - 0.110X_3 - 0.0124X_1X_2 + 0.00842X_1X_3 + 0.00233X_2X_3$ (b) $C_p = 7$, PRESS= 0.429932, $R^2_{pred} = 39.87$, model in problem 20 is better. **25.** (a) $\hat{\beta}_0 = 52.5773, \hat{\beta}_1 = 1.4683, \hat{\beta}_2 = 0.6623$ (b) (94.4596, 97.4732), (90.401, 101.5317). **27.** (a) $\hat{Y} = -0.0784 + 0.000044X_1 + 0.00245X_2 + 0.0183X_3 + 0.00779X_4 - 0.00313X_5$ (b) Fitted model is significant, p-value= 0.000. **29.** (a) $\hat{Y} = 0.012468 - 0.00001X_1X_2 + 0.01988X_2X_3$ (b) $R^2_{pred} = 91.55\%$.
31. (a) $Var(\hat{Y}) = 0.000676265$ (b) $\hat{\sigma}_{\hat{\beta}_1} = 0.001415$, $\hat{\sigma}_{\hat{\beta}_2} = 0.07201$ (c) (-0.007246 ± 0.002928), (1.58934 ± 0.14899). **33.** $\hat{\eta} = -53.0928 + 0.114493X_1 + 1.13665X_2 - 2.44767X_3$. **35.** (a) $\hat{\eta} = -6.61146 + 0.497121X_1 + 0.0507921X_2$ (b) p-values= 0.568, 0.015, β_2 is not significant, β_1 is significant.

37. (a) Males: $\hat{Y} = -97.205 + 0.0415982X_1 + 12.2375X_2 + 3.92213X_3$, females: $\hat{Y} = -86.8758 + 0.0415982X_1 + 12.2375X_2 + 3.92213X_3$. (b) Fitted model is significant at 5% level. (66.1566, 128.270), (30.7276, 163.699). **39.** (a) (71.3893, 115.419) (b) (33.3880, 153.420).

CHAPTER 17

Section 17.2

5. No, Maximum number of linearly independent estimable functions is 2.

Section 17.3

1. Reject H_0, p-value $= 0.000$. **3.** p-value$=$ 0.025, reject H_0 (b) $\hat{\delta}_1 = -1.475$, $\hat{\delta}_2 = 2.125$, $\hat{\delta}_3 = -0.650$. **5.** p-value $= 0.147 > 0.10 = \alpha$, do not reject hypothesis of equal means. **7.** $\delta_2 - \delta_1$: (0.763, 6.737), $\delta_3 - \delta_1$: (− 1.491, 4.041), $\delta_4 - \delta_1$: (3.905, 9.155), $\delta_3 - \delta_2$: (− 5.462, 0.512), $\delta_4 - \delta_2$: (− 0.076, 5.636), $\delta_4 - \delta_3$: (2.63, 7.88). **9.** Do not reject hypothesis of equal means. **11.** $\delta_2 - \delta_1$: (−0.878, 8.078), $\delta_3 - \delta_1$: (−3.653, 5.303), $\delta_3 - \delta_2$: (−7.253, 1.703), yes. **13.** $\delta_2 - \delta_1$: (− 1.359, 4.693), $\delta_3 - \delta_1$: (− 0.644, 5.644), $\delta_4 - \delta_1$: (− 2.444, 3.844), $\delta_5 - \delta_1$: (− 2.080, 5.080), $\delta_3 - \delta_2$: (− 2.005, 3.671), $\delta_4 - \delta_2$: (− 3.895, 1.871), $\delta_5 - \delta_2$: (− 3.481, 3.147), $\delta_4 - \delta_3$: (− 4.764, 1.164), $\delta_5 - \delta_3$: (− 4.423, 2.423), $\delta_5 - \delta_4$: (− 2.623, 4.223), do not reject H_0. Here, we note that all the confidence intervals contain zero. Hence, we may conclude that based on these data all training programs are equally good. Yes.

Section 17.4

1. p-value for coatings and blocks >0.05. **3.** Chemicals are significantly different but machines are not at 10% level. **5.** p-value$=$ 0.083, chemicals are significantly different at $\alpha = 0.10$. **7.** p-value$=$ 0.083, chemicals are not significantly different at $\alpha = 0.05$.

Section 17.5

1. $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, i = 1, 2, 3; j = 1, 2, 3, 4; k = 1, 2;$ $\sum_{i=1}^{3} \alpha_i = 0, \sum_{j=1}^{4}$ $\beta_j = 0$, $\sum_{i=1}^{3} \gamma_{ij} = \sum_{j=1}^{4} \gamma_{ij} = 0$; $\varepsilon_{ijk} \sim N(0, \sigma^2)$, and ε_{ijk}'s are independent.

Source	*DF*	*SS*	*MS*	*F*-ratio	*P*-value
Temperatures	2	68.583	34.2917	3.04	0.086
Raw material	3	36.458	12.1528	1.08	0.396
Interaction	6	167.417	27.9028	2.47	0.086
Error	12	135.500	11.2917		
Total	23	407.958			

3. H_0: $\alpha_1 = \alpha_2 = \alpha_3$ versus H_1: all α_i are not equal, do not reject H_0. **5.** All p-values >0.05. **7.** $P_{\text{int.}} = 0.2616 > 0.05, P_{\text{temp}} = 0.5176 > 0.05, P_{\text{current}} = 0.6832 > 0.05, P_{bl} = 0.7849 > 0.05$. **9.** (b) $P_{\text{int}} = 0.320 > 0.05$ (c) $F_{\text{cloth}} = 7.94$, $F_{\text{mach.}} = 2.68$, $P_{\text{cloth}} = 0.002$, $P_{\text{mach.}} = 0.026$.

Section 17.6

Source	*DF*	*SS*	*MS*	*F*-ratio	*P*-value
Temperature	3	69.25	23.08	5.3303	0.0396
Catalyst	3	14.25	4.75	1.0970	0.4201
Reaction time	3	36.25	12.08	2.7898	0.1318
Error	6	26.0	4.33		
Total	15	145.75			

3. All p-values >0.05.

Section 17.7

1. A: H_0: $\alpha_i = 0$, $MS_A/MS_{AB} \sim F_{3,12}$; B: H_0: $\sigma_\beta^2 = 0$, $MS_B/MS_E \sim F_{4,40}$; AB: H_0: $\sigma_\gamma^2 = 0$, $MS_{AB}/MS_E \sim F_{12,40}$. **3.** p-values for FWA$=$ 0.148, for rolls$=$ 0.294. **5.** p-values for manufacturer$=$ 0.484, for clinics$=$ 0.163. **7.** $p_{Instructor} = 0.902 > 0.01$, $p_{students} = 0.360 > 0.01$.

Review Practice Problems

1. Reject H_0, p-value $= 0.000$. $\delta_2 - \delta_1$: $(-14.20,\ 71.80)$, $\delta_3 - \delta_1$: $(-21.40,\ 64.60)$, $\delta_4 - \delta_1$: $(24.80, 110.80), \delta_3 - \delta_2$: $(-50.20,\ 35.80)$, $\delta_4 - \delta_2$: $(-4.00,\ 82.00)$, $\delta_4 - \delta_3$: $(3.20,\ 89.20)$. **3.** In Problem 2, machine effects not significant. **5.** (a) Catalytic effects significant, p-value $= 0.015$ (b) $\hat{\delta}_I = -1.80, \hat{\delta}_{II} = 0.10, \hat{\delta}_{III} = 1.37$. **7.** (a) p-value $= 0.132$, (b) do not reject H_0. **9.** $t = -2.24$, p-value $= 0.086$. **11.** p-value $= 0.167$, do not reject H_0 **13.** $p_{bl} = 0.006, p_{treat.} = 0.907$. **15.** $p_{Disease} = 0.215 > 0.05, p_{Group} = 0.000 < 0.05$. **17.** (b) $p_{anal.} = 0.178 > 0.05$ (c) $p_{Batch} = 0.000 < 0.05$, $\hat{\beta}_1 = 3.9611$, $\hat{\beta}_2 = -1.3389$, $\hat{\beta}_3 = 3.1056$, $\hat{\beta}_4 = 1.6611$, $\hat{\beta}_5 = -0.0389$, $\hat{\beta}_6 = -1.1389$ (d) A contrast: $\beta_a + \beta_b + \beta_c - \beta_d - \beta_e - \beta_f$, CI $(-1.50,\ -0.44)$, suppliers I has lower effect. **19.** (a) $p_{Analyst} = 0.184 > 0.05$, (b) $p_{metal} = 0.000 < 0.05$, metal effects are significant, $\hat{\beta}_1 = -1.95$, $\hat{\beta}_2 = -1.675$, $\hat{\beta}_3 = -0.425$, $\hat{\beta}_4 = 2.550$, $\hat{\beta}_5 = 1.500$. **21.** (b) $p_{Interaction} = 0.078 > 0.05$, (c) $p_{Poison} = 0.000 < 0.05$, $p_{Treatment} = 0.000 < 0.05$. **23.** (b) $p_{Interaction} = 0.002 < 0.05$, interaction effects are significant. **25.** $p_{Block} = 0.107 > 0.05$, $p_{Interaction} = 0.781 > 05$, $p_A = 0.000 < 0.05$, $p_B = 0.000 < 0.05$, effects of factors A and B are significant. **27.** (b). $p_{Drivers} = 0.005 < 0.05$, driver effects significant, $\hat{\delta}_a = -0.10$, $\hat{\delta}_b = 0.48$, $\hat{\delta}_c = -0.40$, $\hat{\delta}_d = 0.02$ (c) Since $p_{Vehicle} = 0.698$, effects of the vehicles are not significant. **31.** $p_{Machines} = 0.000 < 0.01$, machine effects are significant, $p_{Sample} = 0.259 > 0.01$. **33.** $p_{\Pr ocess} = 0.281 > 0.01$, $p_{Batch} = 0.000 < 0.01$. **35.** (b) $p_{\text{anesthesia}} = 0.896 > 0.05$, (c) Since $p_{\text{mouse}} = 0.991 > 0.05$ do not reject the null hypothesis of no mouse variation within anesthesia. **37.** $p_{Age\ Group} = 0.001 < 0.05$, $\hat{\alpha}_1 = 0.8$, $\hat{\alpha}_2 = 7.1$, $\hat{\alpha}_3 = -2.8$, $\hat{\alpha}_4 = -5.1$. **39.** $H = 7.47, DF = 3$, p-value $= 0.058$ do not reject the null hypothesis of equal means among preparations. **41.** $S = 3.96, DF = 3$, p-value $= 0.266$. **43.** $S = 3.90$, $DF = 3$, p-value $= 0.272$.

CHAPTER 18

Section 18.2

3. $L_1 = Y_1 - Y_2 - Y_3 + Y_4$, $L_2 = -Y_1 + Y_2 - Y_3 + Y_4$, $L_3 = -Y_1 - Y_2 + Y_3 + Y_4$. **5.** Test statistics is $\frac{L}{S\sqrt{\Sigma c_i^2}} = \frac{L}{S\sqrt{4}} = \frac{L}{2S}$, rejection region is $\left|\frac{L}{2S}\right| > t_{n-1,\alpha/2}$. **7.** Answer varies from field to field.

Section 18.3

3. 6, 4, 1. **7.** $\hat{\sigma}/\sqrt{8} = \sqrt{MSE/8}$. **9.** $((\bar{Y}_+^{(A)} - \bar{Y}_-^{(A)}) \pm 0.989\sqrt{MSE}), ((\bar{Y}_+^{(B)} - \bar{Y}_-^{(B)}) \pm 0.989\sqrt{MSE}), ((\bar{Y}_+^{(C)} - \bar{Y}_-^{(C)}) \pm 0.989\sqrt{MSE})$.

Section 18.4

1. (a) A (b) (8.812 ± 1.96) (c) Normal probability plot shows effects A, B, C, AC, BC, and ACD significant at 5% level (d) (8.812 ± 0.7587). **3.** (a) 1.5104 (b) [(Estimates of factor effect) ± 3.202]. **5.** (a) 3.062 (b) 0.8750 (c) B: $(11.1250 \pm 4.029), C$: $(7.8750 \pm 4.029), D$: (5.8750 ± 4.029).

Section 18.5

1. (a) Block 1 (principal block): $1, ab, ac, bc, ad, bd, cd, abcd$; Block 2: $a, b, c, d, abc, abd, acd, bcd$. **3.** (1) $(-,-)$ $1, ad, bc, abcd, abe, ace, cde, bde$, (2) $(+,-)$ $a, d, abc, bcd, be, abde, ce, acde$, (3) $(-,+)$ $b, abd, c, acd, ae, de, abce, bcde$, (4) $(+,+)$ $ab, ac, bd, cd, e, ade, bce, abcde$. **5.** Effects A, B, C, AB, AC, BC, and ACD are significant at the 5% level of significance. **7.** Effects A, B, C, AC. BC, and ACD are significant at the 5% level of significance.

Section 18.6

1. Main effect A significant at the 5% level of significance. **3.** None of the effects is significant at the 5% level of significance. **5.** Only main effect D is significant at the 5% level of significance. **7.** None of the effects is significant at the 5% level of significance. **9.** Interactions AB, BC, and ABC are not significant.

Review Practice Problems

1. (a) $Y_{ijk} = \mu + \sum_i \alpha_i x_i + \sum_{i<j}^{i} \sum_j \alpha_{ij} x_i x_j + \beta_k + \varepsilon_{ijk}; i = 1, 2; j = 1, \ldots, 5; \varepsilon_{ij} \sim N(0, \sigma^2)$ (b) $F_{treat} = 149.10 > 3.4903$ (c) $F_{batch} = 7.80 > 3.2592$, (d) Effect: pH $=$ 51.8, $temp. = 20.2$, (e) $SS_{pH} = 13416.2$, $SS_{temp} = 2040.2$ (f) $\hat{\sigma}^2 = 28.8$, $(-6.43, 4.03)$. **3.** (c) Blocks are significantly different, p-value$=$ 0.000 (d) $\tau = \beta_1 + \beta_2 - 2\beta_3, F_\tau = 56.07 >$ 4.6001 (e) (0.2463, 0.3587). **5.** (a) None of the effects is significant (b) $\hat{\sigma}^2 = 0.228942$ (c) $p -$ value $= 0.594$. **7.** Alias structure is: I + ABCDEF, A + BCDEF, B + ACDEF, C + ABDEF, D + ABCEF, E + ABCDF, F + ABCDE, AB + CDEF, AC + BDEF, AD + BCEF, AE + BCDF, AF + BCDE, BC + ADEF, BD + ACEF, BE + ACDF, BF + ACDE, CD + ABEF, CE + ABDFCF + ABDF, DE + ABCF, DF + ABCE, EF + ABCD, ABC + DEF, ABD + CEF, ABE + CDF, ABF + CDE, ACD + BEF, ACE + BDF, ACF + BDE, ADE + BCF, ADF + BCE, AEF + BCD.

9. None of the main effects or two factor interactions is significant at the 5% level of significance (assuming BC, BD, and CD negligible), since all the corresponding p-values are greater than 0.05, the level of significance.

11. (a) A and AC are significant at $\alpha = 0.05$ (b) No degree of freedom for estimating σ^2. **13.** None of the effects is significant. **15.** (a) M, N, NP, NK, MPK are significant at $\alpha = 0.05$, $\hat{\sigma}^2 = 28.625$. **17.** Effects M, N, NK, ME, NK, MNF are significant at $\alpha = 0.05$. **19.** $A = 1.50$, $B = -25.50$, $C = 27.50$, $D = 38.50$, $AB = -9.50$, $AC = -32.50$, $AD = -17.50$, none of the effects is significant. **23.** $A = 0.075$, $B = 1.625$, $C = -0.425$, $D = 2.925$, $AB = -0.525$, $AC = 0.325$, $AD = 0.175$, No. **25.** (a) $T = -1.12$, $C = 7.37$, $S = -1.87$, $M = -27.62$, $TC = 1.88$, $TS = 4.13$, $TM = -3.13$, $CS = -3.88$, $CM = 4.87$, $SM = -1.37$, $TCS = -4.37$, $TCM = -1.12$, $TSM = 2.13$, $CSM = 1.62$, $TCSM = -3.37$ (b) Main effects C and M are significant at 5% level.

CHAPTER 19

Section 19.2

1. (a) $\hat{Y} = 50.8583 - 0.125x_1 + 1.25x_2 + 1.55x_3$ (b) $MS_{LF} = 26.9873, MS_{PE} = 7.3892$ (c) $F_{LF} = MS_{LF}/MS_{PE} = 3.65$, p-value $= 0.158 > 0.05$. **3.** (a) $\hat{Y} = 25.25 + 1.375x_1 + 2.125x_2 + 1.125x_3 - 0.125X_4$ (b) $MS_{LF} = 1.1875, MS_{PE} = 12.6667$ (c) $F_{LF} = MS_{LF}/MS_{PE} = 0.09$, p-value $= 0.978 > 0.05$, no evidence of lack-of-fit. **5.** (a) $\hat{Y} = 24.875 + 1.375x_1 + 2.125x_2 + 1.125x_3 - 0.125x_4$ (b) $\beta_{12} = \beta_{34}$, $\beta_{13} = \beta_{24}$, $\beta_{14} = \beta_{23}$ (d) Main effects A, B, C are significant at 5% level.

Section 19.3

1. (a) $\hat{Y} = 32.25 + 2.268A + 0.530B - 0.188A * A + 0.062B * B + 5.00A * B$ (b) $\hat{\sigma}^2 = 3.250$ (c) the interaction term is significant at any level greater than the 0.01. **3.** (a) $\hat{Y} = 35.00 + 0.50x_1 + 0.75x_2 - 0.25x_3$ (b) $MS_{LF} = 127.4, MS_{PE} = 70.667$ (c) p-value $= 0.333 > 0.05$, the first-order model is declared good. **5.** (a) $\hat{Y} = 31.98 + 1.47x_1 - 0.97x_2 - 1.45x_3 - 2.00x_1^2 - 2.67x_2^2 - 0.83x_3^2 + 2.25x_1x_2 - 4.00x_1x_3 - 1.00x_2x_3$ (b) $\hat{\sigma}^2 = 16.792$ (c) Both the first and second order terms are not significant.

Section 19.4

1. (a) $\hat{Y} = 32.02 - 2.57w_1^2 + 2.44w_2^2$, saddle. **3.** (a) $\hat{Y} = 35.76 + 4.73w_1^2 - 3.91w_2^2 - 1.63w_3^2$, saddle, minimax point. **5.** (a) $\hat{Y} = 28.75 + 3.00x_1 - 3.00x_2 - 2.75x_3$ (b) $(1, -1, -0.917)(2, -2, -1.834)(3, -3, -2.751)(4, -4, -3.668)(5, -5, -4.585)$.

Review Practice Problems

1. $(0, 0, 0)$, 50.8583; $(0.1, -1, -1.24)$ 47.6738; $(0.2, -2, -2.48)$ 44.4893 $(0.3, -3, -3.72)$ 41.3048; $(0.4, -4, -4.96)$ 38.1203; $(0.5, -5, -6.20)$ 34.9358. **3.** p-value for lack of fit $= 0.004$, not adequate (c) $\hat{Y} = 94.500 + 2.488X_1 + 3.622X_2 - 6.313X_1 * X_1 - 5.563X_2 * X_2 + 5.500X_1 * X_2$ (d) Maximum point, surface is a mound **5.** (b) p-value of lack of fit $= 0.07$, adequate (c) $\hat{Y} = 56.0615 + 3.9125x_1 + 1.8375x_2 - 4.3904x_1^2 + 0.7096x_2^2 - 12.2500x_1x_2$. **7.** $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ biased estimators of β_1, β_2 and β_3, with bias $-\beta_{23}$, $-\beta_{13}$, and $-\beta_{12}$, respectively. **9.** (a) $\hat{Y} = 4.231 - 0.846x_1 - 0.164x_2 - 0.089x_3$ (b) $\hat{\sigma}^2 = 0.01287$. **11.** $F = MS_{LF}/MS_{PF} = 3.05 < 3.48166$, do not reject. **15.** $(1, -0.55, 0.95)$ $(2, -1.10, 1.90)$ $(3, -1.65, 2.85)$ $(4, -2.20, 3.80)$ $(5, -2.75, 4.75)$ $(6, -3.30, 5.70)$. **17.** To estimate the error variance, pure error, and lack-of-fit to determine model adequacy. **19.** (a) $\hat{Y} = 61.10 + 7.875x_1 + 2.625x_2 + 1.375x_3$ (b) Adequate (c) None of the regression coefficients is significant at the 5% level. **21.** (b) $X'X = 8I_5$ (c) Variance optimal $X'X = NI_p = 8I_5$. **23.** (b) $X'X = 16I_5$ (c) Variance optimal $X'X = NI_p = 16I_5$.

CHAPTER 20

Section 20.3 and 20.4

7. (i) 0.9332 (ii) 0.7734 (iii) 0.5000 (iv) 0.0668 (v) 0.0013. **9.** 44, 15, 6, 2, 1.

Section 20.5

1. (a) $\hat{\sigma} = 5.417$ (b) LCL$=$ 33.2298, CL$=$ 40.50, UCL$=$ 47.7702 (c) LCL$=$ 0, CL $=$ 12.6, UCL$=$ 26.6364. **3.** (a) $\hat{\sigma} = 6.945$, (b) LCL =49.8253, CL =60.25, UCL$=$ 70.6747 (c) LCL$=$ 0, CL$=$ 14.3, UCL$=$ 32.6326. **5.** (a) $\hat{\mu} = 20.066$, $\hat{\sigma} = 0.7021$ (b) 17.79%. **7.** UCL $_{\bar{X}} = 20.9460$, CL$_{\bar{X}} = 20.2274$, LCL$_{\bar{X}} = 19.5087$, UCL$_{\bar{S}} = 1.0520$, CL$_{\bar{S}} = 0.5036$, LCL$_{\bar{S}} = 0.00$, the process is in statistical control. **9.** All the values of $\bar{X}$ and S fall within their control limits, so the process in statistical control. **11.** In $\bar{X}$ Control Chart, there are two points that fall outside the control limits, so that the process is not in statistical control. In order to check whether or not the process is behaving according to the specifications, one must first bring the process under statistical control. **13.** The process is in statistical control.

Section 20.6

3. Process is in control. **5.** (a) UCL $= 74.76$, CL $= 52.93$, LCL $= 31.11$ (b) Process is in control. **7.** Point #7 is beyond UCL; revised control limits are UCL$=$ 0.1429, CL$=$ 0.0676, LCL$=$ 0. **9.** $\hat{\lambda} = 8.46$, UCL $= 17.18$, CL $= 8.46$, LCL $= 0$, process is in control. **11.** $\bar{u} = 9.95$, UCL $= 18.69$, CL $= 9.95$, LCL $= 1.21$, process is in control. **13.** ARL$=$ 21, (a) 38 (b) 20 (c) 12.

Section 20.7

1. $\hat{C}_p = 0.8306$, no. **3.** $\hat{C}_p = 1.0374$; process is capable. **5.** Since the value of $\hat{C}_p$ is not sensitive to the location of center of the process, the value of $\hat{C}_p$ does not change.

Review Practice Problems

1. The process is in control. **3.** 1.28%. **5.** (a) $\hat{\mu} = 10.019, \hat{\sigma} = 0.5330$ (b) (8.974, 11.064). **7.** 370. **9.** 0.2222. **11.** (a)UCL $= 16.539$, CL $= 15$, LCL $= 13.461$; UCL$_{\bar{R}} = 5.638$, CL$_{\bar{R}} = 2.667$, LCL$_{\bar{R}} = 0.00$ (b) $\hat{\mu} = 15$, $\hat{\sigma} = 1.1465$ (c) $\hat{C}_p = 1.0176 > 1$; process is capable. **13.** UCL $= 16.6886$, CL $= 15$, LCL $= 13.3114$; UCL$_{\bar{S}} = 2.4719$, CL$_{\bar{S}} =$ 1.1833, LCL$_{\bar{S}} = 0.00$. **15.** UCL$_{\bar{X}} = 2.3076$, CL$_{\bar{X}} = 2.0128$, LCL$_{\bar{X}} = 1.7180$, UCL$_{\bar{S}} =$ 0.4123, CL$_{\bar{S}} = 0.1757$, LCL$_{\bar{S}} = 0$. **17.** 0.223. **19.** (a) UCL$=$ 0.08250, CL$=$ 0.04063, LCL$=$ 0 (b) Process is in control, control limits remains the same. **21.** (a) UCL =0.1014, CL $=0.0487$, LCL $=0$ (b) Process is in control, control limits remains the same. **23.** Again, the process is in statistical control. **25.** (a) np control chart is appropriate, control limits are UCL $=32.55$, CL $=19.45$, LCL $=6.35$ (b) Process is in control, control limits remains the same. **27.** (a) $\hat{p}=0.0333$, (b) UCL $=6.17$, CL $=2$, LCL $=0$, (c) 0.027. **29.** UCL $=0.0298$, CL $=0.0171$, LCL $=0.0044$. **31.** UCL $=31.40$, CL $=18.5$, LCL $=5.60$. **33.** CL $=3.721$, control limits vary from sample to sample. **35.** UCL $=13.3541$, CL $=10$, LCL $=6.6459$.

CHAPTER 21

Section 21.2

3. The CUSUM Control Charts plot the cumulative sum of the deviations of sample values (or sample means, if the sample size is greater than one) from the target value. **5.** $S_i^+=\text{Max}(0, Z_i-k+S_{i-1}^+)$ or $S_i^-=\text{Min}(0, Z_i+k+S_{i-1}^-)$, $i=1,2,3,\dots,10$, $k=\frac{|\mu_1-\mu_0|}{2\sigma}$.

Section 21.3

1. $k=0.5$, $h=5$, process is out of control on the lower side. **3.** CUSUM-chart $k=0.5$, $h=4.77$, process is out of control on lower side at sample 3 and on upper side at sample 13. **5.** $k=0.5$, $h=5$, $(S_0^+=h/2=2.5,\ S_0^-=-h/2=-2.5)$, process is out of control on lower side at sample 1. **7.** Process is in statistical control. **9.** Process is in statistical control.

Section 21.4

1. MA-chart, CL $=11.44$, UCL $=17.92$, LCL $=4.96$, process is in statistical control. **3.** MA-chart, CL $=35.93$, UCL $=37.102$, LCL $=34.758$, process is in statistical control. **5.** MA-chart, CL $=35.93$, UCL $=36.838$, LCL $=35.022$, process is in statistical control.

Section 21.5

1. EMWA-chart, CL $=12$, UCL $=15.752$, LCL $=8.248$, process is in statistical control. **3.** EMWA-chart, $\lambda=0.2$, CL $=11.44$, UCL $=15.76$, LCL $=7.12$, process is in statistical control; EMWA-chart, $\lambda=0.25$, CL $=11.44$, UCL $=16.34$, LCL $=6.54$, process is in statistical control; EMWA-chart, $\lambda=0.3$, CL $=11.44$, UCL $=16.89$, LCL $=5.99$, process is in statistical control. **5.** EMWA-chart, CL $=36$, UCL $=36.3826$, LCL $=35.6174$, process is in statistical control. **7.** EMWA-chart, CL $=36$, UCL $=36.2332$, LCL $=35.7668$, process is in statistical control.

Review Practice Problems

1. $\bar{X}$–R chart, CL $=12$, UCL $=12.3$, LCL $=11.7$; CL $=0.412$, UCL $=0.939$, LCL $=0$, process is out of control; CUSUM-chart, CL $=0$, UCL $=5$, LCL $=-5$, process is out of control at sample 12. **3.** MA-chart, CL $=12$, UCL $=12.268$, LCL $=11.732$, process is out of control at sample 15. **5.** EMWA-chart, $\lambda = 0.4$, CL$=$ 25, UCL$=$ 29.58, LCL$=$ 20.42, process is in statistical control. **7.** CUSUM-chart, $k = 0.5, h = \pm 5$, process is in statistical control. **9.** MA-chart, CL$=$ 24.21, UCL$=$ 26.82, LCL$=$ 21.60, process is out of statistical control at sample 1. **11.** CUSUM-chart, $FIR = h/2$, process is in statistical control. **13.** MA-chart, $m = 3$, CL$=$ 46, UCL$=$ 58.64, LCL$=$ 33.36, $m = 4$, CL$=$ 46, UCL$=$ 56.94, LCL$=$ 35.06, $m = 5$, CL$=$ 46, UCL$=$ 55.79, LCL$=$ 36.21, process is in statistical control. **15.** EMWA-chart $L = 2.5$, CL$=$ 46, UCL =54.03, LCL$=$ 37.97, $L = 3$, CL$=$ 46, UCL = 55.64, LCL$=$ 36.36,$L = 3.5$, CL$=$ 46, UCL =57.24, LCL$=$ 34.76, process is in statistical control. **17.** CUSUM-chart, $FIR = 3h/4$, process is in statistical control. **19.** CUSUM-chart, $FIR = h/2$, process is in statistical control. **21.** EMWA-chart, $L = 2.5$, CL = 6.7685, UCL = 7.0004, LCL = 6.5366, $L = 3$, CL = 6.7685, UCL = 7.0468, LCL = 6.4902, $L = 3.5$, CL = 6.7685, UCL = 7.0931, LCL = 6.4439, process is in statistical control. **23.** CUSUM-chart $k = 0.5$, $h = 4.77$, process is in statistical control. **25.** MA-chart, $m = 3$, CL = 22, UCL = 23.910, LCL = 20.090, process is out of control on the lower side at sample 8; $m = 5$, CL = 22, UCL = 23.480, LCL = 20.520 process came very close to be out of control at samples 3 and 17.

C

BIBLIOGRAPHY

Abonyi, J., and B. Feil (2007). *Cluster Analysis for Data Mining and System Identification*, Springer Science & Business Media.

Agresti, A. (2002). *Categorical Data Analysis*, Wiley.

Akaike, H. (1973). “Information theory and an extension of the maximum likelihood principle.” In *Second International Symposium on Information Theory*, 267–281.

American Society for Testing Materials (1947). *Manual on Presentation of Data*, Philadelphia.

Anderson, E. (1935). “The irises of the Gaspe Peninsula.” *Bulletin of the American Iris society*, 59, 2–5.

Anderson, T.W., and D.A. Darling (1952). “Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes.” *The Annals of Mathematical Statistics*, 23(2), 193–212.

Arnholt, A.T., and B. Evans (2017). “BSDA: Basic Statistics and Data Analysis,” *R package version 1.2.0.* https://CRAN.R-project.org/package=BSDA.

Banfield, J.D., and A.E. Raftery (1993). “Model-based Gaussian and non-Gaussian clustering.” *Biometrics*, 49, 803–821.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). “Fitting linear mixed-effects models using lme4.” *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

Bayes, T. (1763). “An essay towards solving a problem in the Doctrine of chances,” Philosophical Transactions of the Royal Society of London. Reprinted in *Biometrika*, 45 (1958), 296–315.

Bennett, C.A., and N.L. Franklin (1954). *Statistical Analysis in Chemistry and the Chemical Industry*, Wiley.

Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilita*, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8, 3–62.

Bowker, A.H., and G.J. Lieberman (1959). *Engineering Statistics*, Prentice-Hall.

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP, Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

Box, G.E.P. (1952). “Multifactor designs of the first order.” *Biometrika*, 39, 49–57.

Box, G.E.P. (1954). “The exploration and exploitation of response surfaces.” *Biometrics*, 10, 16–60.

Box, G.E.P., and D.R. Cox (1964). “An analysis of transformations.” *Journal of Royal Statistical Society B*, 26, 211–243.

Box, G.E.P., and N.R. Draper (1987). *Empirical Model Building and Response Surfaces*, Wiley.

Box, G.E.P., and J.S. Hunter (1957). “Multi-factor experimental designs for exploring response surfaces.” *Annals of Mathematical Statistics*, 28, 195–241.

Box, G.E.P., and J.S. Hunter (1961). “The 2^{k-p} fractional factorial designs.” *Technometrics*, 3, 311–351, 449–458.

Box, G.E.P., and K.G. Wilson (1951). “On the experimental attainment of optimum conditions.” *Journal of the Royal Statistical Society B*, 13, 1–45.

Box, G.E.P., W.G. Hunter, and J.S. Hunter (1978). *Statistics for Experimenters*, Wiley.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*, Belmont, California: Wadsworth.

Brockmeyer, E., H.L. Halström, and A. Jensen (1948). *The Life and Works of A. K. Erlang*, *Transactions of the Danish Academy of Technical Sciences* (2), 190–192.

Brownlee, K.A. (1960). *Statistical Theory and Methodology in Science and Engineering*, Wiley.

Bulmer, M.G. (1957). “Approximate confidence limits for components of variance.” *Biometrika*, 44, 159–167.

Bustamante, P., Hinkley, D. V., Martin, A., and Shi, S. (1991) “Statistical analysis of the extended Hansen method using the bootstrap technique.” *Journal of pharmaceutical sciences*, 80 (10), pp. 971–977.

Camm, J.D., Cochran, J.J., Fry, M.J., Ohlmann, J.W., and Anderson, D.R. (2014). *Essentials of Business Analytics (Book Only)*, Nelson Education.

Champ, C.M., and W.H. Woodall (1987). “Exact results for Shewhart control charts with supplementary run rule.” *Technometrics*, 29(4), 393–399.

Charytanowicz, M., J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, and S. Zak (2010). “A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images,” *Information Technologies in Biomedicine*, Ed. E. Pietka, and J. Kawa, pp. 15–24. Berlin, Heidelberg: Springer-Verlag.

Chemical Corps Engineering Agency (1953). *Master Sampling Plans for Single, Duplicate, Double and Multiple Sampling*, U. S. Army, Manual No. 2, Army Chemical Center, Maryland.

Cochran, W.G. (1977). *Sampling Techniques*, Wiley.

Cochran, W. G. (1952). “The χ^2 test of goodness of fit.” *The Annals of Mathematical Statistics*, 315–345.

Cochran, W.G., and G.M. Cox (1957). *Experimental Designs*, Wiley.

Conover, W.J. (1999). *Practical Nonparametric Statistics*, Wiley.

Crow, E. L., davis, F. A., and Maxfield, M. W. (1955). *Statistics Manual With Examples Taken From Ordnance Development* (No. Nots-948), Naval ordnance test station china lake, Calif.

Cramer, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press.

Crowder, S.V. (1987). "Computation of ARL for combined individual measurements and moving range charts." *Journal of Quality Technology*, 19(1), 98–102.

Crowder, S.V. (1989). "Design of exponentially weighted moving average schemes." *Journal of Quality Technology*, 21(2), 155–162.

Dalal, S.R., E.B., Fowlkes, and B. Hoadley (1989). "Risk analysis of the space shuttle: pre-challenger prediction of failure." *Journal of the American Statistical Association*, 84(408), 945–957.

Daniel, C. (1959). "Use of half-normal plots in interpreting two-level experiments." *Technometrics*, 1, 311–342.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, Wiley.

Daniel, W.W. (1990). *Applied Nonparametric Statistics*, Duxbury Press.

Daniel, W.W. (2018). *Biostatistics, a Foundation for Analysis in the Health Sciences*, Wiley.

Davies, O.L. (1954). *Designs and Analysis of Industrial Experiments*, Oliver and Boyd.

Dean, A., and D. Voss (1999). *Design and Analysis of Experiments*, Springer-Verlag.

DeLury, D.B. (1950). *Values of the Integrals of the Orthogonal Polynomials*, Toronto, Ontario, Canada: University of Toronto Press.

Deming, W.E. (1986). *Out of the Crisis*, MIT Press.

Dixon, W.J., and F.J. Massey, Jr. (1957). *Introduction to Statistical Analysis*, McGraw-Hill.

Dodge, H.F., and H.G. Romig (1959). *Sampling Inspection Tables* (2nd edition), Wiley.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 45–70.

Draper, N.R., and H. Smith (1981). *Applied Regression Friedman Analysis*, Wiley.

Duncan, A.G. (1958). *Quality Control and Industrial Statistics*, R.D. Irwin, Inc.

Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, (5th edition), Irwin, Homewood, IL.

Dunnett, C.W. (1964). "New tables for multiple comparisons with a control." *Biometrics*, 3, 1–21.

Epskamp, S., A.O.J. Cramer, L.J. Waldorp, V.D. Schmittmann, and D. Borsboom (2012). "qgraph: network visualizations of relationships in psychometric data." *Journal of Statistical Software*, 48(4), 1–18. http://www.jstatsoft.org/v48/i04/.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications* (2nd edition), Wiley.

Fisher, R.A. (1925). “Applications of ‘Student’s’ distribution.” *Metron*, 5, 90–104.

Fisher, R.A. (1936). “The use of multiple measurements in taxonomic problems.” *Annals of Eugenics*, 7(2), 179–188.

Fraser, D.A.S. (1958). *Statistics: An Introduction*, Wiley.

Freund, J.F. (2004). *Modern Elementary Statistics*, Pearson/Prentice Hall.

Freund, R.J., and W.J. Wilson (1998). *Regression Analysis*, Academic Press.

Friedman, M., and L.J. Savage (1947). “Planning experiments seeking maxima,” *Techniques of Statistical Analysis*, Eds. C. Eisenhart, M.W. Hastay, and W.A. Wallis, McGraw-Hill, pp. 365–372.

Furth, J., A.C. Upton, and Kimball, A.W. (1959). “Late pathologic effects of an atomic detonation and their pathogenesis.” *Radiation Research Supplement*, 1, 243–264.

Gayton, W.F., W. Steinroeder, C. Bonnica, and A.C. Loignon (2009). “An investigation of home disadvantage in Davis Cup tennis.” *Perceptual and Motor Skills*, 109, 382–386.

Govindarajulu, Z. (1999). *Elements of Sampling Theory and Methods*, Prentice Hall.

Gumbel, E.J. (1958). *Statistics of Extremes*, Columbia University Press.

Gupta, B.C., and H. Fred Walker (2005). *Applied Statistics for the Six-Sigma Green Belt*, Milwaukee: ASQ Quality Press.

Gupta, B.C., and H. Fred Walker (2007). *Statistical Quality Control for the Six-Sigma Green Belt*, Milwaukee: ASQ Quality Press.

Haddow, J.E., Neveux, L.M., Palomaki, G.E., Lambert-Messerlian, G., Canick, J.A., Grenache, D.G., Lu, J (2011). “*The relationship between PTH and 25-hydroxy vitamin D in pregnancy*,” *Clin Endocrinol (Oxf)*, 75(3):309–314, PMID: 21521334.

Hahsler, M., and M. Piekenbrock (2018). “dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms,” *R package version 1.1-3.* https://CRAN.R-project.org/ package=dbscan.

Hald, A. (1952). *Statistical Theory with Engineering Applications*, Wiley.

Halperin, M. (1952). “Maximum likelihood estimation in truncated samples.” *The Annals of Mathematical Statistics*, 23(2), 226–238.

Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*, Elsevier.

Hawkins, D. M. (1981). “A CUSUM for a scale parameter.” *Journal of Quality Technology*, 13(4), 228–231.

Hawkins, D.M., and D.H. Olwell (1998). *Cumulative Sum Charts and Charting for Quality Improvement*, Springer-Verlag.

Hawkins, D. M. (1993). “Cumulative sum control charting: an underutilized SPC tool.” *Quality Engineering*, 5(3), 463–477.

Hennig, C. (2018). “fpc: Flexible Procedures for Clustering,” *R package version 2.1-11.* https://CRAN.R-project.org/package=fpc.

Hicks, C.R. (1982). *Fundamental Concepts in the Design of Experiments*, Holt, Rinehart and Winston.

Hinkelmann, K., and O. Kempthorne (2005). *Design and Analysis of Experiments*, Vol. 2, Wiley.

Hoel, P.G. (1954). *Introduction to Mathematical Statistics* (2nd edition), Wiley.

Hogg, R.V., and A.T. Craig (1959). *Introduction to Mathematical Statistics*, Macmillan.

Hogg, R.V., and J. Ledolter (1992). *Applied Statistics for Engineers and Physical Scientists*, Macmillan.

Hogg, R.V., and E.A. Tanis (1993). *Probability and Statistical Inference*, Macmillan.

Hothorn, T., and Zeileis, A. (2015). “partykit: A modular toolkit for recursive partytioning in R.” *The Journal of Machine Learning Research*, 16(1), 3905–3909.

Hothorn, T., K. Hornik, and A. Zeileis (2006). “Unbiased recursive partitioning: a conditional inference framework.” *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

Huber, P.J. (2011). *Robust Statistics*, Berlin: Springer-Verlag.

Ishikawa, K. (1976). *Guide to Quality Control*, Asian Productivity Organization.

Jacobson, J. (1998). “St. Luke’s Hospital Breaks Out of the Pack to Improve Patient and Physician Satisfaction,” Milwaukee: ASQ. To access this case study, visit http://asq.org/knowledge-center.

Jacobson, J. (2008). “St. Luke’s Hospital Breaks out of the Pack to Improve Patient and Physician Satisfaction Pool of Case Studies,” Milwaukee: *The American Society for Quality* (*ASQ*).

Jayalath, K.P. (2018). “A machine learning approach to analyze ODI cricket predictors.” *Journal of Sports Analytics*, 4(1), 73–84.

Jayalath, K.P., and R.F. Gunst (2017). “The maximum likelihood based intensity estimate for circular point processes.” *Environmental and Ecological Statistics*, 24(3), 449–468.

Johnson, R.A. (2000). *Miller and Freund’s Probability and Statistics for Engineers*, Prentice Hall.

Johnson, R.A., and D.W. Wichern (2007). *Applied Multivariate Statistical Analysis*, Upper Saddle River, New Jersey: Pearson Prentice Hall International, Inc.

Juran, J.M. (1974). *Quality Control Handbook*, McGraw-Hill.

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*, Wiley.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*, Wiley.

Kenney, J.F., and E.S. Keeping (1956). *Mathematics of Statistics*, Vols. 1 and 2 (3rd edition), Van Nostrand.

Khuri, A.I., and J.A. Cornell (1996). *Response Surfaces Design and Analysis*, Dekker.

Kimball, A.W. (1960). "Estimation of morality intensities in animal experiments," *Biometrics*, 16, 505–521.

Kruskal, W. H., and Wallis, W. A. (1952). "Use of ranks in one-criterion variance analysis." *Journal of the American statistical Association*, 47(260), 583–621.

Kutner, M.H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*, McGraw-Hill.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, Wiley.

Lee, E.T., and J.W. Wang (2003). *Statistical Methods for Survival Data Analysis*, Wiley.

Ligges, U., and M. Mächler (2003). "Scatterplot3d - an R package for visualizing multivariate data." *Journal of Statistical Software*, 8(11), 1–20.

Lohr, S.L. (1999). *Sampling Design and Analysis*, Duxbury Press.

Lucas, J.M. (1982). "Combined Shewhart CUSUM quality control scheme." *Journal of Quality Technology*, 14(2), 51–59.

Lucas, J.M., and R.B. Crosier (1982). "Fast initial response for CUSUM quality control scheme." *Technometrics*, 24(3), 199–205.

Lucas, J.M., and M.S. Saccucci (1990). "Exponentially weighted moving average control schemes: properties and enhancements." *Technometrics*, 32(1), 1–29.

Lucas, J.M., D.J. Davis, and E. Saniga (2006). "Detecting improvements using Shewhart attribute control charts when the lower control limit is zero." *IIE Transactions*, 38, 659–669.

Mann, N.R., R.E. Schafer, and N.D. Singpurwalla (1974). *Methods for Statistical Analysis of Reliability and Life Data*, Wiley.

McDonald, G.C., L.C., Vance, and D.I. Gibbons (1995). "Some Tests for Discrimination Between Lognormal and Weibull Distributions: An application to Emissions Data," *Recent Advances in Life Testing and Reliability*, Ed. N. Balakrishnan, pp. 475–487. CRC Press.

McNeil, D.R. (1977). *Interactive Data Analysis*. New York: Wiley.

Melnykov, V., and R. Maitra (2010). "Finite mixture models and model-based clustering." *Statistics Surveys*, 4, 80–116.

Mendenhall, W.M. and T.L. Sincich (2007). *Statistics for Engineers and the Sciences*, Pearson/Prentice Hall.

Miller, I., and M. Miller (2004). *John E. Freund's Mathematical Statistics with Applications*, Pearson/Prentice Hall.

Millikan, R.A. (1930). "Values of the electron and related constants." *Physical Review*, 35, 1231.

Mitchell, T.M. (1997). *Machine Learning*, New York: McGraw-Hill.

Montgomery, D.C. (2009a). *Design and Analysis of Experiments*, Wiley.

Montgomery, D.C. (2008). *Introduction to Statistical Quality Control*, Wiley.

Montgomery, D.C. (2011). *Applied Statistics and Probability for Engineers*, Wiley.

Montgomery, D.C., E.A. Peck, and G.G. Vining (2006). *Introduction to Linear Regression Analysis*, Wiley.

Mood, A.M., and Graybill, F.A. (1963). *Introduction to the Theory of Statistics*, McGraw-Hill.

Murphy, R.B. (1948). "Non-parametric tolerance limits." *Annals of Mathematical Statistics*, 19, 581–589.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*, Duxbury Press.

Myers, R.H., and D.C. Montgomery (1995). *Response Surface Methodology, Process and Product Optimization Using Designed Experiments*, Wiley.

Myers Raymond, H., and Montgomery, D. C. (2002). *Response Surface Methodology: Process and Product Optimization using Designed Experiment*, A Wiley-Interscience Publication.

Nativ, O., Y. Raz, H.Z. Winkler, Y. Hosaka, E.T. Boyle, T.M. Therneau, and M.M. Lieber (1988). "Prognostic value of flow cytometric nuclear DNA analysis in stage C prostate carcinoma." *Surgical Forum*, (Vol. 39, pp. 685–687). American College of Surgeons.

Nelson, W.B., and N. Doganaksoy, (1995). "Statistical Analysis of Life or Strength Data from Specimens of Various Sizes Using the Power-(log) Normal Model," *In Recent Advances in Life Testing and Reliability*, Ed. N. Balakrishnan, pp. 378–408. CRC Press.

NIST and SEMATECH (2003). *Engineering Statistics Handbook*, National Institute of Standards and Technology (NIST). http://www.itl.nist.gov/div898 /handbook/.

Parzen, E. (1960). *Modern Probability Theory and Its Applications*, Wiley.

Penrose, K.W., A.G. Nelson, and A.G. Fisher (1985). "Generalized body composition prediction equation for men using simple measurement techniques." *Medicine and Science in Sports and Exercise*, 17(2), 189.

Quesenberry, C.P. (1997). *SPC Methods for Quality Improvement*, Wiley.

Quinlan, J.R. (1986). "Induction of decision trees." *Machine Learning*, 1, 81–106.

Raiffa, H., and R. Schlaifer (1961). *Applied Statistical Decision Theory*, Graduate School of Business Administration, Harvard University.

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

Rawlings, J.O. (1988). *Applied Regression Analysis, A Research Tool*, Wadsworth.

Reid, C. (1982). *Neyman from Life*, New York: Springer-Verlag.

Reinsel, D., J. Gantz, and R. Rydning (2018). Data Age 2025, The Digitization of the World From Edge to Core, *An IDC White Paper*.

Resnikoff, G.J., and G.J. Lieberman (1957). *Tables of the Non-Central t-Distribution*, Stanford University Press.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge, England: Cambridge University Press.

Roberts, S.W. (1959). "Control chart tests based on geometric moving averages." *Technometrics*, 1(3), 239–250.

Rohatgi, V. (1984). *Statistical Inference*, Wiley.

Ross, S.M. (1996a). *Introduction to Probability and Statistics for Engineers and Scientists*, Academic Press.

Ross, S.M. (1996b). *Introductory Statistics*, McGraw-Hill.

Russell, V.L. (2009). "Response-surface methods in R, using RSM." *Journal of Statistical Software*, 32(7), 1–17. http://www.jstatsoft.org/v32/i07/.

Ryan, T.P. (2000). *Statistical Methods for Quality Improvement*, Wiley.

Saniga, E.M., D.J. Davis, and J.M. Lucas (2009). "Using Shewhart and CUSUM Charts for Diagnosis with Count Data in Vendor Certification Study." *Journal of Quality Technology*, 41(3), 217–227.

Schafft, H.A., T.C. Staton, J. Mandel, and J.D. Shott (1987). "Reproduceability of electro-migration measurements." *IEEE Transactions on Electron Devices*, ED-34, 673–681.

Scheaffer, R.L., W. Mendenhall III, and R.L. Ott (2006). *Elementary Survey Sampling*, Duxbury: Thomson, Brooks/Cole.

Scheffe, H. (1953). "A method for judging all contrasts in the analysis of variance." *Biometrika*, 40, 87–104.

Scheffe, H. (1959). *The Analysis of Variance*, Wiley.

Schwarz, G. (1978). "Estimating the dimensions of a model." *Annals of Statistics*, 6, 461–464.

Scrucca, L. (2004). "qcc: an R package for quality control charting and statistical process control." *R News*, 4/1, 11–17.

Scrucca, L., M. Fop, T.B. Murphy, and A.E. Raftery (2016). "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." *The R Journal*, 8(1), 205–233.

Shewhart, W.A. (1931). *Economic Control of Manufactured Product*, ASQ Press.

Shewhart, W.A. (1939). *Statistical Methods From the Viewpoint of Quality Control*, Courier Devor Publications.

Siegle, S. (1965). *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill.

Snow, G. (2016). "TeachingDemos: Demonstrations for Teaching and Learning," *R package version 2.10.* https://CRAN.R-project.org/package=TeachingDemos.

Somerville, P.N. (1958). "Tables for obtaining non-parametric tolerance limits." *Annals of Mathematical Statistics*, 29, 599–601.

Statistical Research Group, Columbia University (1948). *Sampling Inspection*, McGraw-Hill.

Steel, R.G.D., and J.H. Torrie (1960). *Principles and Procedures of Statistics*, McGraw-Hill.

Stephane, C. (2017). “pwr: Basic Functions for Power Analysis,” *R package version 1.2-1.* https://CRAN.R-project.org/package=pwr.

Stephen, M. (2018). “rpart.plot: Plot ‘rpart’ Models: An Enhanced Version of ‘plot.rpart’,” *R package version 3.0.5.* https://CRAN.R-project.org/package=rpart.plot.

Stone, G.C., and J.F. Lawless (1979). “The applications of Weibull statistics to insulation aging test.” *IEEE Transactions on Electrical Insulation*, EI-14, 233–239.

Student (1908). “On the probable error of the mean.” *Biometrika*, 6, 1–25.

Sukhatme, P.V., and B.V. Sukhatme (1970). *Sampling Theory of Survey with Applications* (2nd edition), Ames: Iowa State University Press.

Therneau, T. (2017). “A Package for Survival Analysis in S. version 2.38. 2015,” *Reference Source.* http://CRAN.R-project.org/package=survival.

Therneau, T.M., and E.J. Atkinson (1997). “An introduction to recursive partitioning using the RPART routines.”

Therneau, T., and B. Atkinson (2018). “rpart: Recursive Partitioning and Regression Trees,” *R package version 4.1-13.* https://CRAN.R-project.org/package=rpart.

Therneau, T.M. and P.M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*, New York: Springer. ISBN: 0-387-98784-3.

Tufte, E.R. (2001). *The Visual Display of Quantitative Information*, Vol. 2, Cheshire, CT: Graphics Press.

Tukey, J.W. (1953). *The Problem of Multiple Comparisons*, unpublished notes, Princeton University.

Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison Wesley.

United States Department of Defense (1950). *Military Standard. Sampling Procedures and Tables for Inspection by Attributes* (Military Standard 105A.), Government Printing Office.

Van Dobben de Bruyn, C. S. (1968). Cumulative sum tests: Theory and practice.

Venables, B. (2013). “conf.design: Construction of factorial designs,” *R package version 2.0.0.* https://CRAN.R-project.org/package=conf.design.

Venables, W.N., and B.D. Ripley (2002). *Modern Applied Statistics with S* (4th edition), New York: Springer. ISBN: 0-387-95457-0.

Wackerly, D., Mendenhall, W. and Scheaffer, R.L. (2008). *Mathematical Statistics with Applications*, Brooks/Cole.

Wald, A. (1947). *Sequential Analysis*, Wiley.

Walpole, R.E., Myers, R.M. and Myers, S.L. (2007). *Probability and Statistics for Engineers and Scientists*, Pearson/Prentice Hall.

Weibull, W. (1951). "Statistical Distribution Function of Wide Applicable Applicability, ASME Journal of Applied Mechanics, Transactions of the American Society of Machanical Engineers" *Journal of Applied Mechanics, Transactions ASME*, 18(3), 293–927.

Weksi, B. (2018). "kmed: Distance-Based k-Medoids," *R package version 0.1.0.* https://CRAN.R-project.org/package=kmed.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer-Verlag.

Wickham, H., M. Lawrence, D.T. Lang, and D.F. Swayne (2008). "An introduction to rggobi." *R News*, 8(2), 3–7.

Western Electric, (1985). *Statistical Quality Control Handbook*, Western Electric Corporation, Indianapolis, IN.

Wilks, S.S. (1948). *Elementary Statistical Analysis*, Princeton University Press.

Western Electric, (1956). *Statistical Quality Control Handbook*, Western Electric Corporation, Indianapolis, IN.

Wilks, S.S. (1962). *Mathematical Statistics*, Wiley.

Woodward, R.H., and P.L. Goldsmith (1964). *Cumulative Sum Techniques*, Oliver and Boyd.

Wu, S.M., (1964). "Tool life testing by response surface methodology, I and II," *Transactions of the ASME, Journal of Engineering for Industry*, 86, 105.

Yates, F. (1958). *The Design and Analysis of Factorial Experiments*, Commonwealth Agriculture Bureau.

Index

Statistics and Probability with Applications for Engineers and Scientists using MINITAB, R and JMP, Second Edition. Bhisham C. Gupta, Irwin Guttman, and Kalanka P. Jayalath.

Companion website: www.wiley.com/college/gupta/statistics2e

WILEY END USER LICENSE AGREEMENT